공학석사 학위논문

영상 기반 실내 위치 추정을 위한 심층 합성곱 신경망에 관한 연구

- 전방향 영상 데이터 정제 및 전이학습 방법론 비교 -

A Study on Image Based Deep Convolutional Neural Network for Indoor Positioning

- Method of Data Refinement with Omnidirectional Image and Comparison of Transfer Learning Methodologies -

2019년 8월

서울대학교 대학원

건설환경공학부

김 광 중

영상 기반 실내 위치 추정을 위한 심층 합성곱 신경망에 관한 연구

- 전방향 영상 데이터 정제 및 전이학습 방법론 비교 -

지도교수 김 용 일

이 논문을 공학석사 학위논문으로 제출함
2019년 5월

서울대학교 대학원
건설환경공학부
김 광 중

김광중의 공학석사 학위논문을 인준함
2019년 7월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

국문초록

위치 기반 서비스 산업은 다양한 산업에서의 IoT(Internet of Things), 고속 통신 네트워크 확산, 실시간 위치 추적 및 교통정보, 위치 기반 검색 및 광고·마케팅 등으로 그 규모와 중요성이 증대되고 있다. 위치 기반 서비스에서 위치 추정 기술은 알맞은 서비스 제공을 위한 기본 자료로써 필수적인 역할을 한다. 따라서 위치 기반 서비스의 품질 향상을 위해 정확한 위치 추정 기술의 필요성이 대두되고 있으나, 범용적으로 사용되는 위치 추정 기술인 위성 항법 장치는 실내에서 비교적 정확도가 낮아 다른 형태의 기술들이 요구된다. 이중에서, 스마트폰의 보급으로 스마트폰에 탑재된 센서들을 이용한 실내 위치 추정 방법이 범용성면에서 뛰어나고, 특히 탑재된 카메라로 촬영한 영상을 분류하여 위치를 추정하는 방식은 추가적인 장비 설치가 필요없어 경제적이며, 통신에 독립적이기 때문에 안정적으로 작동할 수 있다는 강점을 가진다. 또한 최근 합성곱 신경망을 이용한 영상의 분류 기술이 큰 발전을 이루어 이를 이용한 실내 위치 추정 시스템이 주목받고 있다.

하지만 좋은 성능을 발휘하는 합성곱 신경망을 온전히 처음부터 학습시키기 위해서는 방대한 학습 데이터의 수량, 계산 자원과 계산 시간이 필요하다. 따라서 대부분의 경우 사전 학습된 합성곱 신경망을 사용하는 전이학습을 사용한다. 전이학습을 이용하여 합성곱 신경망을 학습하기 위해서는 다음과 같은 요인을 고려해야 한다.

먼저 전이학습에 사용할 심층 합성곱 신경망의 종류를 결정해야 한다. 2012년 AlexNet이 출현한 이후, 여러 기법을 활용하여 효율성과 정확도 면에서 발전된 다수의 심층 합성곱 신경망이 개발되었기 때문에, 전이학습에 이용할 합성곱 신경망간의 실내 위치 추정 성능을 비교하는 연구가 필요하다. 다음으로 전이학습 유형을 선택해야한다. 전이학습은 사전 학습 데이터와 대상 데이터간의 유사도, 대상 데이터 수량에 따라 네 가지 유형으로 나눌 수 있다. 그러므로 심층 합성곱 신경망의 전이학습 시, 전이학습 유형을 설정하고, 설정한 전이학습 유형 간 성능을 비교하여야

한다. 또한 대량의 데이터를 얻는 방법을 고려해야 한다. 전이학습을 이용하여도 선행 연구들에서 구역별 학습 영상의 수는 평균 1,000장 이상이 사용된 바 있다. 그러므로 대량의 영상을 효율적으로 취득하는 방법을 고안할 필요가 있다. 마지막으로 정확도 저하를 유발하는 데이터를 제거하는 방법이 연구되어야 한다. 대량으로 영상 취득 시, 신경망 학습에 악영향을 끼치는 영상이 포함될 수 있다. 이들을 제거하는 데이터 정제가 대량의 영상을 얻는 방법에 뒷받침 되어야한다.

본 연구는 심층 합성곱 신경망을 통한 영상 분류로 실내 위치 추정 시, 심층 합성곱 신경망의 종류와 전이학습 유형에 따른 성능을 비교하고, 대량의 영상을 효율적으로 취득하기 위해 전방향 영상을 이용하는 방법을 제안한다. 또한 전방향 영상 사용 시, 학습에 악영향을 끼치는 영상을 제거하는 데이터 정제를 제안한다.

본 연구에서는 개발된 신경망 중 AlexNet, MobilNet V2, 그리고 Inception-ResNet V2를 실험에 사용하였다. 위 신경망은 모두 ILSVRC(ImageNet Large Scale Visual Recognition Competition)라는 대회에서 사용된 영상 데이터베이스인 ImageNet의 영상 분류를 위한 신경망으로, Inception-ResNet V2, MobileNet V2, AlexNet 순으로 ImageNet의 영상 분류 정확도가 높다고 알려져 있다(Mathwork, 2019).

본 연구의 실험은 가상공간에서 실행되었으므로, 실험을 위한 가상공간 구축방법을 설명하였다. 현실의 상황에 가정하여 장소별로 공통점과 특징을 부여하였고, 전방향 카메라의 자세와 경로를 설정하였다. 단위 구(unit sphere)에 색상정보가 매핑 된 전방향 영상에서 다수의 핀홀 카메라 모델을 생성하여 원근 투영 영상으로 분할하고, 변환하는 방법을 제안하였다. 제안한 방법으로 1장의 전방향 영상에서 30장의 원근투영 영상을 생성하였다.

생성된 원근 투영 영상을 이용하여 ImageNet의 영상으로 사전 학습된 AlexNet, MobilNet V2, 그리고 Inception-ResNet V2를 학습시켰다. 이때 전이학습 유형별 분석을 위해 각각의 신경망에 대해 전역 연결 계층만 학습시키는 유형1과 합성곱 계층도 부분적으로 학습시키는 유형3으로

경우를 나누어 학습시켜 총 여섯 경우로 학습을 진행하였다.

데이터 정제는 영상의 정보량을 계산하기 위한 엔트로피와 정제되지 않은 영상으로 학습된 신경망을 이용하였다. 엔트로피가 낮은 순으로 정렬하고 영상이 정분류 된 빈도가 오분류 된 빈도보다 커질 때 까지 제거할 영상의 엔트로피 기준을 증가시켰다.

실험 결과 ImageNet의 영상 분류 정확도가 높은 순서대로 실내 위치 추정 정확도가 높았기 때문에 ImageNet의 영상 분류 정확도가 높은 신경망을 사용하는 것이 정확도 면에서 실내 위치 추정에 유리함을 확인하였다. 또한 정확도는 모든 신경망에서 전이학습 유형1보다 유형3으로 학습시킨 경우가 평균적으로 6.12% 높았으므로 유형3이 유형1보다 실내 위치 추정에 유리함을 확인하였다. 마지막으로, 데이터 정제 후 학습시킨 신경망은 정제하지 않은 데이터로 학습시킨 신경망보다 평균 1.99%의 정확도가 향상되었다.

본 연구의 결과로 초기에 개발된 신경망을 사용하기보다는 발전된 신경망을 이용하고, 그리고 전역 연결층만을 학습시키기보다는 특징 추출을 위한 계층을 부분적으로 학습시키는 것이 실내 위치 추정에서 더 정확한 결과를 기대할 수 있음을 실험적으로 확인했다는 점에서 의의가 있다. 또한 제안된 전방향 영상을 이용하여 대량의 원근 투영 영상을 생성하는 방법은 합성곱 신경망의 학습에서 데이터 수량 부족 문제를 해소할 수 있다는 점에서 가치가 있다. 그리고 학습에 비효율적인 영상을 제거하는 데이터 정제는 대량의 영상을 얻는 방법에 범용적으로 활용될 수 있다. 본 연구의 결과는 합성곱 신경망을 이용한 실내 위치 추정 연구들에서 활용되고, 전이학습을 이용한 영상 기반 합성곱 신경망은 실내 위치 추정의 수단 중 하나로 적용이 가능할 것으로 판단된다.

주요어 : 실내 위치 추정, 심층 합성곱 신경망, 전방향 영상, 데이터 정제

학 번 : 2017-22313

목 차

표 목 차

그 림 목 차

1. 서 론

1.1 연구배경

ICT(Information and Communication Technology) 기술이 급속하게 발달되고 있음에 따라, 이들을 기반으로 한 다양한 서비스 및 시스템들의 융합이 빠르게 진행되고 있다(유재준, 2013). ICT 기반 융합 흐름의 한 가지는 위치정보를 이용한 위치 기반 서비스 산업이다. 위치 기반 서비스 산업은 다양한 산업에서의 IoT 도입, 고속 통신 네트워크 확산, 실시간 위치 추적 및 교통정보, 위치 기반 검색 및 광고·마케팅 등으로 인해 규모와 중요성이 증대되고 있다. 위치 기반 서비스를 제공하기 위해 위치 추정 기술은 기본 자료로써 필수적인 역할을 한다. 따라서 위치 기반 서비스의 품질을 위해 오차가 적은 정확한 위치 추정 기술의 필요성이 대두되고 있다(한국인터넷진흥원, 2017).

위치 추정에는 주로 범지구적 위성 항법 장치(GNSS(Global Navigation Satellite System))가 사용되나, 이 기술은 건물, 구조물 밀집 지역, 실내와 같은 폐색지역에서 신호의 수신이 어렵다는 한계를 가진다. 그러므로 실내에서는 GNSS 대신 무선신호를 기반으로 하는 RFID(Radio Frequency Identification), Wi-Fi(Wireless Fidelity), Beacon, MEMS(Micro systems) Sensor, Li-Fi(Light Fidelity), Geo-Magnetic UWB(Ultra Wide Band) 등의 기술이 위치 추정에 주로 사용된다(Lowry *et al.*, 2015). 이들은 스마트폰에 내장된 센서들을 이용하여 실내 위치를 추정하는 기술로, 스마트폰의 대중화로 인해 접근성면에서 강점을 가져 활발하게 연구되고 있다. 하지만 위의 무선신호 기반 기술들은 무선신호세기의 불안정, 무선신호를 송·수신하는 액세스 포인트의 공간적 배열의 민감성 등의 이유로 안정적으로 위치를 추정하기 어려울 가능성이 있으며, 비교적 측위 가능한 범위가 제한적이고, 추가적인 장비설치가 필요하다(Werner *et al.*, 2011).

반면 스마트폰 카메라와 같은 광학센서를 통해 얻은 시각정보를 이용하여 위치를 추정하는 방법은 추가적인 설치장비가 필요하지 않아 구축 비용이 적으며, 무선신호를 사용하지 않아 적용범위가 넓고, 통신에 독립적으로 구축 가능하며, 센서가 소형화 됨에도 주변환경의 형상을 정밀하고 풍부하게 얻을 수 있어 최근 활발하게 연구되고 있다(오정현, 2018).

시각정보를 이용한 실내 위치 추정은 SIFT(Scale Invariant Feature Transform)(Lowe, 1999, 2004), SURF(Speeded-Up Robust Features)(Bay *et al.,* 2006, 2008)와 같은 지역 특징 서술자나 전역 특징 서술자를 계산하고 계산된 특징 서술자 기반의 BoW(Bag of Words)알고리즘을 사용하는 연구가 대부분을 차지하였다(Sünderhauf *et al.*, 2015). 하지만 서술자와 BoW 알고리즘을 이용한 방법은 약간의 환경변화나 시점변화가 발생할 경우 성능이 낮아지는 문제가 있다(오정현, 2019). 최근 컴퓨터 비전과 기계학습의 발전으로 심층 합성곱 신경망(Deep Convolutional Neural Networks)을 이용한 방법이 시각적 인식, 분류 그리고 탐지 등의 다양한 분야에서 다른 방법보다 뛰어난 성능을 보이고 있다(Sünderhauf *et al.,* 2015). 하지만 실내 위치 추정 분야에서 심층 합성곱 신경망을 사용한 연구는 아직 2012년에 개발된 AlexNet이나 그 변형, 혹은 더 단순한 구조를 사용한 연구가 대부분이며, AlexNet의 출현 이후 개발된 더 효율적이거나 좋은 성능을 발휘하는 심층 합성곱 신경망 구조를 이용한 장소인식 기반 실내 위치 추정 연구가 부족한 실정이다.

심층 합성곱 신경망과 같은 심층 신경망, 혹은 학습시킬 모수가 많은 기계학습 알고리즘은 일반적으로 사전 정보 없이 원하는 성능을 내도록 학습시키기 위해서 충분한 크기의 학습 데이터가 필요하다. 복잡하고 모수가 많은 신경망이나 알고리즘일수록 학습 데이터가 적으면 학습 데이터에만 지나치게 최적화되는 과적합(overfitting)현상 발생 가능성이 높아진다(Srivastava *et al.*, 2014). 또한 사전정보 없이 초기화 시킨 심층 합성곱 신경망을 학습시키기 위해서는 범용적으로 사용하기 힘든 수준의 계산 자원과 계산 시간이 필요하다. 영상 분류 대회인 ILSVC에서 우승

한 신경망들은 영상 백만개 이상을 이상으로 고성능 그래픽 처리 장치와 며칠간의 학습시간을 통하여 학습되었다(강대기, 2016). 또한 여러 가지 기법들을 적용하고 초모수들을 최적화하는 과정에도 많은 시간이 소모된다. 그러므로 학습데이터 규모, 계산 자원 그리고 시간의 문제로 인하여 심층 합성곱 신경망 활용 시, 초기화 상태의 심층 합성곱 신경망을 처음부터 학습하기보다는 대부분의 경우, 학습된 모수와 신경망 구조를 이용하는 전이학습을 이용한다.

전이학습을 이용했음에도 분류에 사용하는 학습데이터의 개수는 선행연구 분석결과 평균적으로 분류 종별 1,000개 내외가 요구되므로 (Kornblith *et al.*, 2018) 대량의 학습 데이터를 효율적으로 취득하는 방법이 고안되어야한다. 하지만 대량의 학습 영상을 취득할 때 이상치를 보이는 영상이나 신경망에 악영향을 주는 영상이 포함될 수 있다. 그러므로 대량으로 영상을 취득할 경우, 위의 데이터를 제거하는 방법을 고려해야 한다. 또한 영상 분류를 위한 심층 합성곱 신경망은 2012년부터 발전을 계속하여 여러 종류가 존재한다. 그러므로 실내 위치 추정을 위해 전이학습에 사용할 사전 학습된 심층 합성곱 신경망의 종류에 따른 비교분석이 수행되어야 한다. 전이학습 유형도 다수가 존재하므로, 심층 합성곱 신경망을 선택 한 후, 심층 합성곱 신경망의 학습 범위에 따른 성능을 분석하는 전이학습 유형별 분석이 필요하다.

1.2 연구 동향

실내 위치 추정을 위한 시각적 장소인식은 지역 특징 서술자를 이용하는 방법, 전역 특징 서술자를 이용하는 방법이 주로 연구되었으며, 최근에는 심층 학습(deep learning)이 널리 쓰이기 시작하여 심층 학습을 적용하는 연구도 늘어나고 있다(Ali *et al*, 2017).

지역 특징 서술자를 이용한 방법을 통한 위치 추정은 다양한 방식으로 영상의 키포인트(keypoint)를 추출하고, 이들 주변의 특징(feature)을 추출한 후, 이를 이용하여 미리 구축된 장소 별 이미지들과 매칭하는 방식으로 이루어진다. SIFT와 SURF는 시각적 장소 인식에 가장 널리 사용되는 지역 서술자로써, 영상의 회전, 크기 변화, 그리고 조명 변화에 강건하다는 장점이 있으나 연산량이 많다는 단점이 있다. 이후 개발된 다양한 서술자들도 시각적 장소인식에 사용되었다. Harris affine regions(Mikolajczyk *et al.*, 2001, Ho e*t al.*, 2007), BRIEF(Calonder *et al.*, 2010, 2012), BRISK(Leutenegger *et al.*, 2011), ORB(Rublee, 2011), FREAK(Alahi *et al.*, 2012), FAST(Rosten *et al.*, 2006, 2010) 등 다양한 서술자가 장소 인식에 사용된 바 있다. 지역 서술자를 이용하여 영상의 특징들을 추출한 후 영상과 매칭하는 방법으로 초창기에는 특징들을 이용하여 학습 영상과 입력 영상을 직접 비교하는 방식이 사용되었다. 하지만 영상 하나에 있는 모든 특징들을 비교하는 방식은 비효율적이므로 BoW 방법이 영상을 매칭하는 방법으로 제안되었다(Sivic and Zisserman, 2003, Fei-Fei and Perona, 2005). BoW는 학습 영상에서 추출한 특징이 구역별 출현하는 빈도로 사전을 생성하고, 입력 영상에서 특징을 추출해 특징들의 빈도를 비교하여 영상을 분류하는 방법이다. 이 방법은 영상들의 기하관계를 고려하지 않기 때문에 학습 영상과 입력 영상의 시점(viewpoint)이 변화에 강건하다는 장점이 있다. 그러나 약간의 환경 변화에도 키포인트들이 달라져 성능이 매우 떨어진다는 단점 또한 존재한다.

전역 특징 서술자는 영상의 키포인트를 감지하는 단계가 없이 영상의

전체적인 특징을 이용한다. 여기에는 영상 전체에서 SURF 서술자를 추출하는 whole-image SURF(WI-SURF)(Badino *et al.*, 2012)를 이용한 방법과 BRIEF-gist(Oliva and Torralba, 2001)라는 전역 특징 서술자를 이용한 방법이 제안되었고 이들은 약간의 변화에 강건한 장소 인식 방법임이 확인되었으나, 시점변화에는 불리하다는 한계가 있다. 이후 2012년 ILSVC라는 영상 분류 대회에서 심층 학습을 이용한 심층 합성곱 신경망인 AlexNet이 우승하며 영상 분류에 심층 합성곱 신경망이 뛰어남을 증명하였고 이는 실내 위치 추정 연구에도 영향을 끼쳤다.

심층 학습은 신경망에 대해 추상적인 개념을 계층적 구조로 학습시킬 수 있으며, 시각적 물체 인식, 감지, 분류, 그리고 장소 인식 분야에서 다른 방법보다 뛰어난 성능을 보이고 있다(LeCun *et al.*, 2015). 특히 심층 합성곱 신경망은 영상, 동영상, 음성처리 분야의 발전을 이끌고 있다. Sunderhauf는 다양한 장소의 이미지를 합성곱 신경망으로 학습을 수행하고 신경망 중간의 결과를 영상의 서술자로 활용하면 더 높은 성능을 나타내기도 하였으며, 합성곱 신경망의 중반부분의 계층에서 변화와 시점에 강건한 특징들을 얻을 수 있음을 밝혀내었다(Sunderhauf *et al.*. 2015). Zhang은 전이학습을 사용하지 않고 단순한 합성곱 층 4개와 완전연결 계층 1개를 구조로 하는 신경망의 모수를 처음부터 학습시켜 실내 위치 추정에 사용하였다(Zhang *et al.*, 2016). 신경망을 초기화하고 처음부터 학습시켰기 때문에 분류할 장소 당 약 15,000여개의 영상을 학습영상으로 사용하였다. 그러나 초기 특징값 추출을 위한 필터를 보면 잡음이 많은 필터들이 관찰된다. 이는 학습이 완료되지 않았다는 뜻이고 학습영상의 개수가 초기 상태부터 학습시키기에는 다소 부족했기 때문으로 판단된다. Werner는 AlexNet과 AlexNet을 약간 변형한 CaffeNet에서 마지막 전역 연결 계층의 차원을 분류할 장소의 개수로 변경하고 모수와 구조를 그대로 가져오는 전이학습을 이용하여 시각적 장소 인식을 이용한 실내 위치 추정을 실행하였다(Werner *et al.*, 2016). Werner는 분류할 구역마다 약 1,000개의 영상을 학습영상으로 사용하였다. Kim은 실시간 실내 위치 추정을 위해 CaffeNet의 마지막 층의 차원을 변환하고 전체

모수를 가져온 후, 이를 구역 당 5,000여개의 영상을 학습 데이터로 사용하여 작은 학습률로 미세조정(fine tune) 시켰다(Kim and Chen, 2015). Chen과 Ha는 ILSVC 2014년의 준우승 모델인 VGG16를 이용하여 특징값을 추출하고 이를 카메라 자세정보와 결합하여 실내 위치를 추정하였다(Chen *et al.,* 2018)(Ha *et al.*, 2018). 그러나 이들은 초창기 합성곱 신경망인 AlexNet, VGG16을 기반으로 분류하여 위치를 추정하였기 때문에 이 후 등장한 더 좋은 성능을 발휘하는 신경망을 시각적 실내 위치 추정에 접목하여 더욱 개선할 여지가 있다. 또한, 전이학습을 이용한 선행연구의 경우, 수정할 계층을 결정하는 전이학습 유형별 분석이 이루어진 연구가 부족하여, 그 부분에 대한 분석도 필요하다.

데이터 정제는 주로 합성곱 신경망 학습의 사전 준비단계에서 이상치를 제거하기 위한 방법으로 사용되었다. Zhang은 영상을 입력받는 합성곱 신경망을 이용하여 실내 위치 추정 시스템을 구축할 때, 학습영상을 얻기 위한 방법으로 동영상 촬영 후 동영상의 프레임들을 학습영상으로 이용하는 방법을 적용하였다. 이때 움직임으로 인해 흐려진 영상을 제거하는 방법으로 흐림 지수(blur index)를 사용하여 흐린 영상을 탐지 후 제거하였다. Kim, Werner, Chen, Ha의 연구에서도 학습데이터를 취득 후 흐린 영상, 노이즈가 심한 영상을 정제한 바 있다. 그러나 사전 준비단계에서 이상치를 제거하는 방법이외에 데이터를 정제하는 방법을 고려한 연구는 부족한 것으로 조사되었다.

1.3 연구 목적 및 범위

본 연구는 사전 학습 된 심층 신경망을 기반으로 실내 공간에서 영상을 이용하여 사용자 위치를 추정할 때, 대량의 학습 영상 데이터를 효율적으로 취득하기 위해 전방향 카메라(Omnidirectional camera)를 이용한 방법을 소개한다. 또한 얻은 영상 데이터를 바탕으로 전이학습 유형별, 심층 신경망 모델별 실내 위치 추정 정확도와 학습시간을 비교 분석하였으며, 이를 위해 구역별 정확도 분석을 선행하였다. 또한 전방향 카메라를 이용하여 학습 영상 데이터를 취득할 시, 빈 벽, 바닥 등 정보가 없는 영상들이 존재하므로 이 영상들을 제거할 수 있는 데이터 정제를 제안하고 정제한 후의 영상들로 심층 신경망을 학습하였을 때의 결과와 정제하지 않은 영상들로 학습하였을 때의 결과를 비교하였다.

본 논문은 다음과 같은 순서로 구성된다. 먼저 2장에서는 영상 분류에 사용된 심층 신경망의 종류와 특징을 소개하고 사전 학습된 심층 신경망을 이용하여 학습하는 전이학습방법론에 대하여 알아본다. 그리고 3장 실험방법에서는 본 논문의 실험이 이루어진 가상환경과 카메라 경로에 대해 다룬다. 또한 학습 영상으로 사용할 원근 투영 영상들을 전방향 영상에서 분할, 변환하여 구축하는 방법을 제안한다. 그 후 심층 신경망을 학습시킬 때 고려한 점과 방법, 학습 영상을 정제한 방법을 설명한다. 실험의 결과와 분석을 정리한 4장에서는 학습된 심층 합성곱 신경망과 시험 영상을 이용하여 실내 위치 추정 결과를 정확도와 학습 시간, 분류 시간의 관점에서 비교하여 분석한다. 마지막 5장에서는 본 연구가 가지는 의의와 한계점을 언급한다.

2. 심층 합성곱 신경망의 전이학습

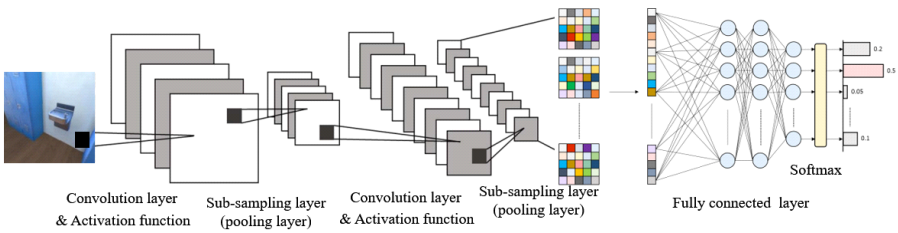
전이학습으로 학습된 심층 합성곱 신경망을 실내 위치 추정에 사용하려면 이용할 심층 합성곱 신경망과 전이학습 유형을 선정하는 과정이 필요하다. 따라서 2장에서는 전이학습에 사용할 심층 합성곱 신경망을 선정하기 위하여 종전까지 개발된 심층 합성곱 신경망의 구조와 종류를 설명하고, 사전 학습된 심층 합성곱 신경망을 사용하는 전이학습 유형들을 다룬다.

2.1 심층 합성곱 신경망

‘2.1 심층 합성곱 신경망’에서는 심층 합성곱 신경망의 구조와 종류로 나누어 설명한다. 먼저 ‘2.1.1 신경망의 구조’에서는 합성곱 신경망을 이루고 있는 각 계층들의 원리와 역할을 설명한다. 그 다음 ‘2.1.2 신경망의 종류’에서는 ImageNet의 영상 분류에서 좋은 성능을 발휘한 대표적인 신경망들의 종류와 특징을 살펴보고 실험에 사용할 신경망들을 선정한다.

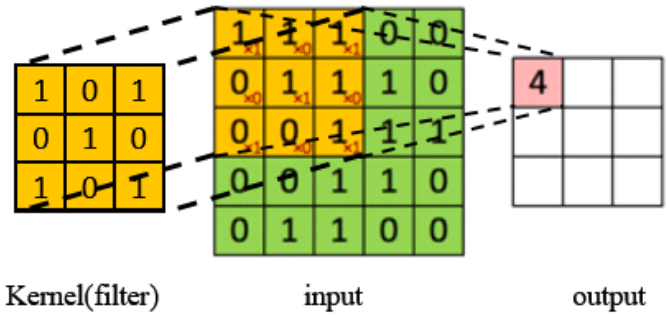
2.1.1 심층 합성곱 신경망 구조

합성곱 신경망에는 일반적으로 합성곱 계층(convolution layer), 활성함수(activation function), 하위 추출 계층(sub-sampling layer), 전역 연결 계층(fully connected layer), 그리고 softmax 함수가 존재한다. [그림 2-1]은 간단한 영상 분류 합성곱 신경망의 구조이다. 영상이 합성곱 신경망의 입력으로 사용되고 출력은 해당 영상이 각 집단마다 속할 확률이다.



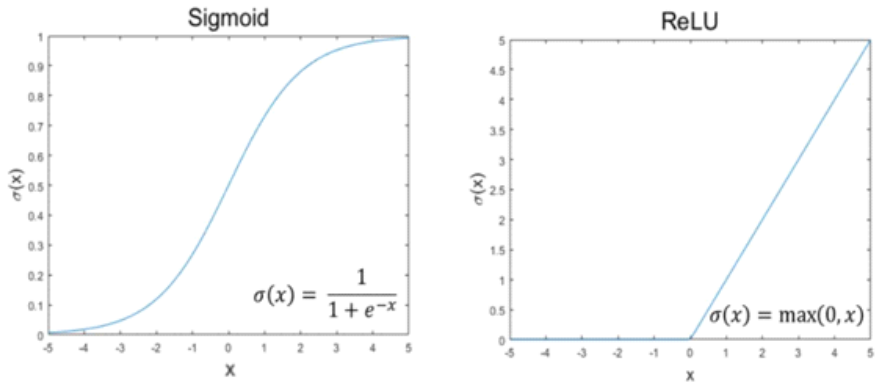
[그림 2-1] 합성곱 신경망의 기본구조
(추민곤, 2019)

합성곱 계층에서는 3차원으로 이루어진 커널(kernel, 혹은 필터(filter))과 입력값의 합성곱 연산으로 입력값인 영상의 특징들을 추출한다. 합성곱은 [그림 2-2]와 같이 주어진 커널을 가중치로 곱한 후 더하는 방식으로 이루어지며 정한 간격만큼 커널을 이동시키며 진행된다.



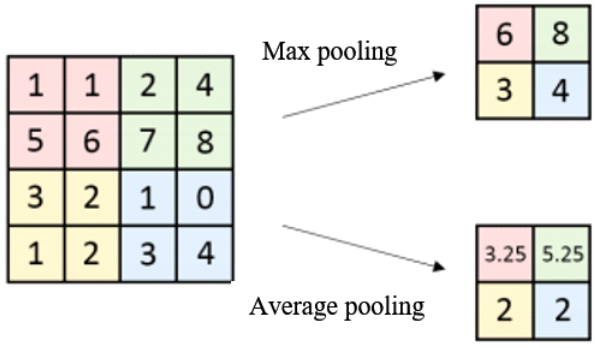
[그림 2-2] 커널과 입력의 합성곱 연산

활성함수는 입력값과 출력값 사이의 비선형성을 추가하기 위하여 사용된다(Specht, 1990). 가장 널리 사용하는 활성함수는 [그림 2-3]의 시그모이드(sigmoid)와 ReLU(Rectified Linear Unit)이다. 시그모이드는 깊지 않은 합성곱 신경망에서 주로 사용되어 왔으나, 최근에는 더 깊은 합성곱 신경망을 구축하며 계산시간을 단축시키기 위해 비교적 간단한 ReLU 함수를 많이 사용하고 있다. 여러 개의 ReLU를 이용하면 비선형 함수를 도출할 수 있고, 시그모이드보다 좋은 효과를 가짐이 입증되었다(Mass *et al.*, 2013).



[그림 2-3] 시그모이드와 ReLU 활성화 함수

하위 추출 계층은 불필요한 값들을 줄이고, 특징들의 크기를 감소시켜 학습 속도를 증가시키는 역할을 한다. 널리 사용하는 하위 추출 기법으로 가장 큰 값만 선택하는 최대값 추출(max-pooling)과 평균값을 계산하여 그 값만을 사용하는 평균값 추출(average-pooling)이 있다. 최대값 추출은 생물의 신경 세포 원리를 모방하여, 범위내의 가장 큰 결과값만을 선택하는 방식이다. 평균값 추출은 특징값의 잡음 제거에 주로 사용되어 완전 연결 계층에 입력되기 전 사용된다. [그림 2-4]는 특징값들을 2×2크기로 하위 추출하는 예시이다.



[그림 2-4] 2×2 최댓값 추출과 평균값 추출

전역 연결 계층은 추출된 특징값들을 이용한 분류기(classifier) 역할을 하고 있다. 종합하면, 전역 연결 계층 이전 신경망 계층 (합성곱 계층과 하위 추출 계층)은 특징 추출기(feature extractor)이며, 전역 연결 계층은 추출된 특징을 이용한 분류기이다. 전역 연결 계층에서 추출된 특징값들에 모수인 가중치를 곱하고 편향을 더한다. 그 후, 전역 연결 계층의 결과 값을 수식(2-1)의 softmax 함수에 입력하여 영상이 집단에 속할 확률로 변환하여 분류에 사용한다.

$$p_k = \frac{\exp(y_k)}{\Sigma_{c=1}^{N}\exp(y_c)} \qquad (2\text{-}1)$$

p_k : 영상이 k 집단에 속할 확률, N : 분류하는 집단의 개수

y_k : 최종 전역 연결 계층에서 k번째 출력값

합성곱 신경망의 학습과정은 다음과 같다. 먼저 데이터를 입력값으로 입력하여 softmax 함수값을 구한 후 손실을 계산한다. 그 후, 손실함수를 바탕으로 마지막 계층부터 손실을 최소화하기 위해 모수를 수정하는 오류 역전파 알고리즘을 사용한다. 전역 연결 계층과 softmax 함수값인 확률값을 기반으로 손실함수를 구할 수 있다. 한 입력 데이터의 손실함수는 수식(2-2)의 cross-entropy 함수를 사용하고 신경망의 손실은 학습, 혹은 검증에 사용하는 모든 데이터의 손실함수 값의 평균으로 계산한다.

$$H_{p_k'} = -\Sigma_{k=1}^{N} p_k' \cdot \log(p_k) \qquad (2\text{-}2)$$

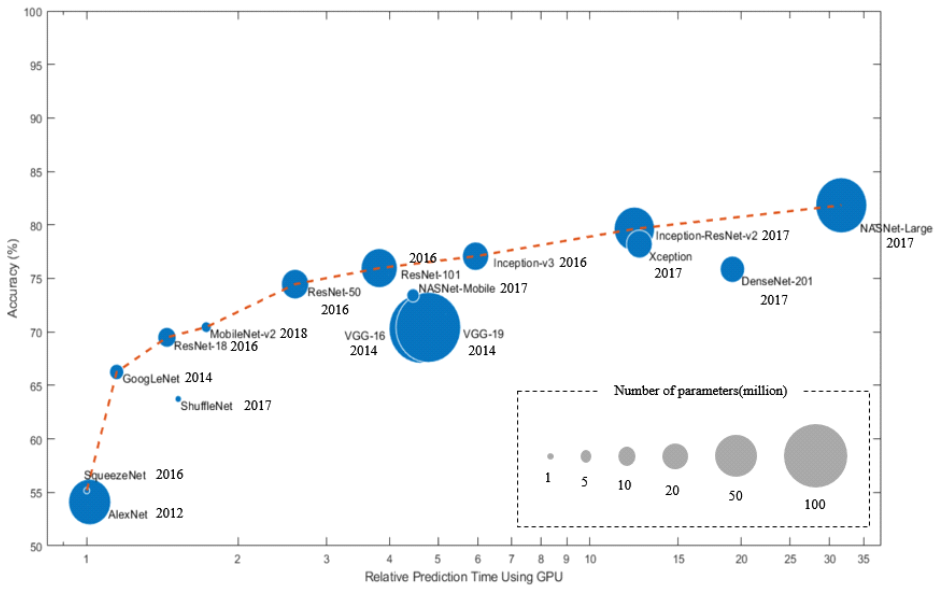
p_k' : 입력 데이터가 c 집단에 속할 때, $k=c$이면 1, 그 외에는 0

정리하자면, 신경망이 학습되는 과정에서 합성곱 계층의 모수와 전역 연결 계층의 모수는 손실을 감소시키는 방향으로 수정되고, 학습완료 후 합성곱 계층에서 커널은 분류에 사용되는 특징을 추출하는 데 최적화된 커널이며, 완전 연결 계층은 주어진 신경망 구조에서 추출된 특징을 이용한 분류에 최적화된 분류기라고 할 수 있다.

신경망을 학습시킬 때, 데이터들을 직접 학습에 사용할 학습 데이터(training data)와 신경망 학습이 잘 진행되는지를 검증하기 위한 검증 데이터(validation data)로 분리하여 학습시킨다. 신경망이 학습됨에 따라 학습 데이터의 손실은 감소하나 검증 데이터의 손실은 증가할 수 있는데 이는 학습 영상군의 특성을 신경망이 과도하게 반영한 과적합이 발생한 것이다. 과적합이 발생하면 학습되지 않은 데이터에 대한 분류 성능이 감소하여 범용성이 떨어지게 된다. 이러한 과적합 현상은 학습 데이터에 비해 모수가 많거나 복잡할 때 발생한다. 이를 해결하기 위해서 합성곱 신경망의 구조 변경, 학습 영상의 개수 증가 그리고 학습을 조기 종료(early stop) 하는 방법이 제안되어 사용되고 있다(Caruana *et al.*, 2001).

2.1.2 영상 분류를 위한 심층 합성곱 신경망 종류

영상 분류를 위한 합성곱 신경망은 2012년 AlexNet의 출현 이후 [그림 2-5]와 같이 다수 출현하였다. [그림 2-5]는 ImageNet의 영상 데이터베이스로 1,000개의 사물을 분류하기 위하여 개발된 합성곱 신경망별 분류 정확도, 영상을 입력하였을 때 예측까지 걸리는 시간, 개발된 연도, 그리고 신경망의 모수 개수를 나타낸 그래프이다. [그림 2-5]의 가로축은 AlexNet을 기준 값 1로 하여 각 신경망 별로 상대적 예측 시간을 나타낸 것이다. 한편 신경망의 학습시간은 주로 손실을 계산하기 위해 학습 영상을 입력하여 예측하는 시간과 검증 영상을 이용하여 예측 후, 검증 정확도 (validation accuracy)를 계산하는 시간의 합으로 결정된다. 그러므로 학습된 신경망을 사용하여 분류를 수행하는데 소요되는 예측시간 역시 학습 시간과 비례하는 경향을 보인다. 따라서 신경망 별 학습 시간 역시 [그림 2-5]의 가로축을 통해 비교할 수 있다. [그림 2-5]의 세로축은 정확하게 분류한 영상의 개수를 전체영상으로 나눈 정확도를 나타낸다. 또한, 원의 크기로 신경망의 모수의 개수를 나타내고, 예측시간이 비슷하며 최고의 정확도를 가지는 신경망들을 점선으로 연결하였다. 또한 개발된 연도를 이름 옆에 기재하였다.



[그림 2-5] 신경망 별 ImageNet의 영상
분류 정확도, 분류 시간, 개발 연도
그리고 모수의 개수를 나타내는 그래프
(MathWorks, 2019의 그림에 연도를 추가함)

[그림 2-5]에서 연도를 보면 심층 합성곱 신경망은 비슷한 정확도를 보이지만 모수와 예측 시간 감소를 목적으로 하는 효율성과 모수와 예측 시간이 증가하더라도 정확성을 증가시키는 방향으로 발전하고 있음을 알 수 있다. 본 논문에서도 효율성과 정확성을 이용하여 실내 위치 추정을 위한 영상 분류에 사용할 심층 합성곱 신경망을 선정하였다. 다음은 위 [그림 2-5]에 표시된 심층 합성곱 신경망 중 영상 분류 신경망에서 중요한 기점이 된 신경망들과 그에 대한 설명이다.

AlexNet(2012)

AlexNet은 종전까지 최고 정확도를 보인 SIFT 기반의 우승 모델보다 약 10%의 정확도 향상을 보이며 2012 ILSVCR 영상 분류 대회에서 우

승한 신경망이다. 이 모델은 영상 분류에 심층 합성곱 신경망이 기존의 기술보다 훨씬 더 좋은 성능을 낼 수 있음을 보여주었다. AlexNet은 5개의 합성곱 계층과 3개의 전역 연결 계층 구조를 통해 전년 대비 10%의 성능을 향상시켰다. AlexNet의 설계자는 학습 단계 전, 학습 영상을 좌우반전, 평행이동, 회전 등과 같은 데이터 증가(data augmentation) 기법을 사용하였다. 또한 과적합 문제를 해결해기 위해 드롭아웃(dropout) 기법을 적용하였다. 드롭아웃은 신경망 학습시 모수를 확률적으로 수정하지 않는 방법으로 학습 영상에만 모수가 반영되는 것을 방지하는 기법이다. AlextNet은 합성곱 계층에서 11×11, 5×5, 3×3 크기의 커널을 사용했으며 2개의 그래픽 처리 장치로 6일 동안 학습시켰다고 한다 (Krizhevsky *et al.*, 2012).

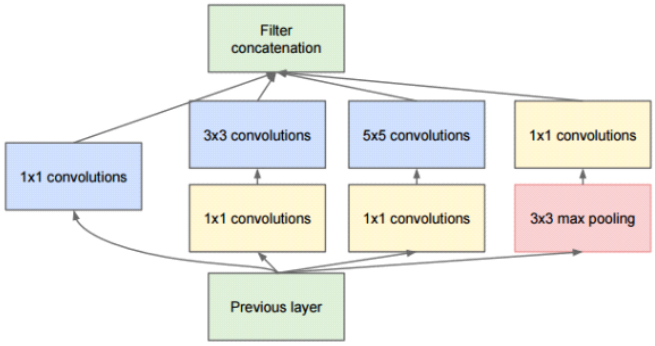
VGGNet(2014)

VGGNet은 AlexNet보다 더 깊은 구조를 가지는 합성곱 신경망이다. VGGNet에 사용된 기본 아이디어는 큰 크기의 커널 한 개보다 3×3 크기의 커널을 사용한 합성곱 계층을 다수 쌓은 경우가 비선형성을 더 잘 설명하고 성능이 좋다는 것으로, AlexNet의 경우는 11×11 크기의 커널을 사용한 합성곱 계층을 사용하였으나, VGGNet의 경우는 3×3 크기의 커널을 사용한 합성곱 계층을 여러 겹 쌓아 11×11 크기의 커널을 가진 합성곱 계층과 유사한 효과를 내지만 모수의 개수를 감소시킬 수 있었다. VGGNet은 단순한 구조이지만 성능이 우수하고 변형이 쉬워 2104년 ILVRC 우승 신경망인 GoogLeNet보다 많이 사용되었고 주로 16층의 VGG16구조와 19층짜리 VGG19구조가 많이 사용되었다. 그러나 VGGNet은 단순한 만큼 계속해서 깊이 쌓을 수 있었으나, 어느 정도 이상의 깊이에서는 학습이 이루어지지 않는 한계가 있다.(Simonyan *et al.,* 2014).

GoogLeNet(2014), Inception-v3(2016)

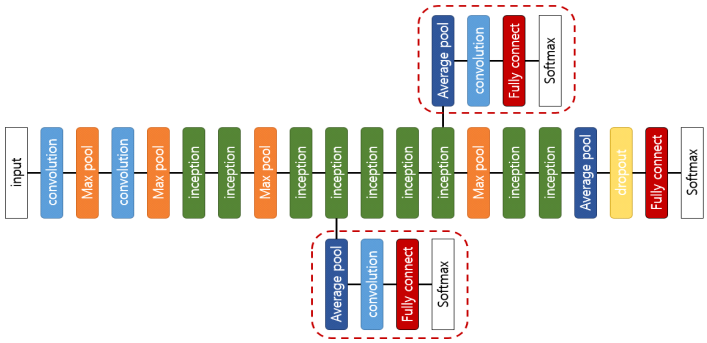
GoogLeNet은 2014년 ILSVRC 영상 분류 대회에서 1등을 수상한 신경망이다. Google에서 inception이라는 이름으로 발표하였으며 inception과 GoogLeNet이라는 이름을 혼용해 사용한다. 당시 심층 신경망의 성능은 구조가 깊어질수록, 즉 합성곱 계층이 많을수록, 그리고 각각의 계층이 넓을수록 성능이 좋다고 알려져 있었다. 하지만 구조가 깊어지고 계층이 넓을수록 모수가 증가하면서 학습단계에서 과적합이 발생할 가능성이 커진다. 또한 학습단계에서 최종 결과값과 가까운 계층의 모수부터 수정되는 오류 역전파 알고리즘을 사용하였으나, 이것은 최종 결과값으로부터 먼 계층들의 모수는 수정이 잘 되지 않는 기울기 소실(Gradient Vanishing)이라는 문제를 야기하였다. GoogLeNet은 위 두 가지 문제의 해결방법을 반영한 하여 다음과 같은 기법을 사용한 신경망이다.

먼저, 모수를 줄이면서 5×5, 7×7커널과 같은 효과를 내도록 작은 커널(3×3, 1×1)을 가진 합성곱 계층 다수를 병렬적으로 계산 후 결과값을 깊이 방향으로 쌓는(concatenation) 형태로 신경망을 구성하였다. VGGNet에서는 단순히 3×3의 계층을 다수 적층하는 방식이었다면 GoogLeNet에서는 [그림 2-6]과 같이 한 계층에서 여러개의 합성곱을 적용하고 합치는 방식을 사용했다고 볼 수 있다. GoogLeNet의 저자는 [그림 2-6]과 같은 모듈을 inception 모듈로 명명하였다.



[그림 2-6] inception 모듈의 구조
(Szegedy *et al.*, 2015)

또한 기울기 소실을 해결하기 위해, 신경망 중간 계층에 최종 결과값을 출력하는 계층을 추가하였다. [그림 2-7]은 GoogLeNet(inception)의 구조이며, 추가된 출력층을 점선으로 표시하였다(Szegedy *et al.*, 2015).



[그림 2-7] GoogLeNet의 구조
(Szegedy *et al.*, 2015)

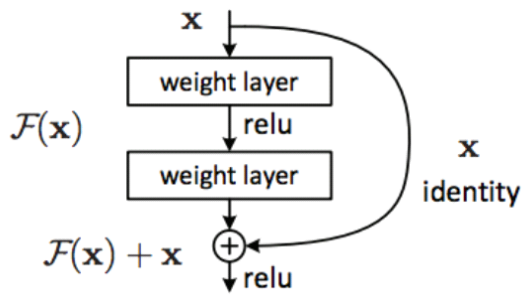
출력층을 중간에 추가하는 기법으로 오류 역전파가 마지막 출력층뿐만 아니라 추가된 중간 출력층에서도 시작되기 때문에 기울기 소실을 감소시켰고 이는 효과적인 학습을 가능하게 하였다.

GoogLeNet이 출현한 이후, 개선 버전으로 inception-v2,v3,v4가 발표되었다. 이중 v3은 배치 정규화라는 신경망 계층의 분포를 정규분포로 변환하는 기법을 채택하여 기울기 소실을 방지하고 학습 속도를 빠르게 하였으며, inception 모듈에서 3×3 커널을 3×1, 1×3 커널 두 개로 변경하는 방법을 사용하여 모수 감소 효과를 극대화 하였다(Szegedy *et al.*, 2016).

ResNet(2015), Inception-ResNet-V2(2017)

ResNet은 2015년 ILSVRC 영상 분류대회를 우승한 신경망으로 Top-5 error 3.57%를 달성하였다. 우승한 모델의 구조는 2014년의 우승 신경망인 VGG16보다 8배 이상 더 깊은 계층인 152계층을 쌓은 구조이다. 깊은 구조의 신경망은 위에서 언급하였듯 모수가 많아지고 학습이 어려워지는 문제가 있다. He는 skip connection이라는 새로운 개념을 제안하여

깊은 구조의 신경망에서 발생하는 문제를 해결하였다. 기존의 신경망은 합성곱 계층의 입력이 반드시 합성곱 계층의 연산을 거치고 다음 계층의 입력으로 사용되었다. 그러나 skip connection을 이용하면 [그림 2-8]과 같이 합성곱 계층의 입력은 합성곱 연산을 거친 후의 결과값과 합성곱 연산을 생략한 입력값을 더한 후 다음 계층의 입력으로 전달된다(He *et al.,* 2016).



[그림 2-8] skip connection의 구조
(He *et al.,* 2016)

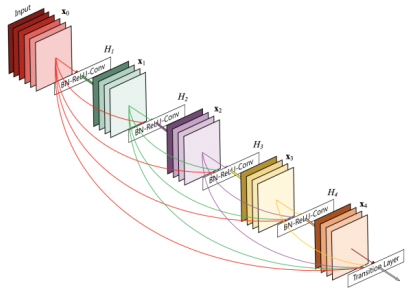
Skip connection의 효과로 신경망 학습 시 다음 계층을 건너뛰면서 입력과 출력이 연결되기 때문에 오류 역전파의 경로가 단순해지는 효과를 얻는다. 결과적으로 더욱 깊은 망도 학습이 가능하였고, 깊어진 만큼 정확도를 개선할 수 있었다.

Inception-ResNet-V2는 2014년의 Inception 신경망에 ResNet의 skip connection의 개념을 추가하여 설계한 신경망이다(Szegedy *et al.*, 2017).

DenseNet(2017)

ResNet 이후 영상 분류를 위하여 개발된 신경망들은 skip connection 만큼의 획기적인 발전없이 ResNet과 Inception을 혼합하여 사용하는 신경망이 대다수였다. 2016 ILSVRC 대회의 우승 신경망도 ResNet과 Inception을 혼합한 신경망으로 큰 주목을 받지 못하였다. 이러한 배경에서 2017년 DenseNet이 발표되었다. ResNet이 skip connection을 다음

계층에만 적용되는 구조인 반면, DenseNet은 skip connection을 [그림 2-9]와 같이 전체 계층에 추가하여 발전시킨 신경망이다.

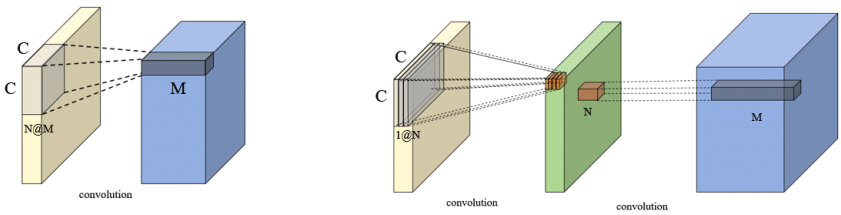


[그림 2-9] DenseNet이 적용한 여러 계층을
건너뛰는 skip connection의 구조
(Huang *et al.*, 2016)

DenseNet의 개발자는 다수의 skip connection으로 인해 각 계층의 특징값 전파를 강화하고 특징값을 재사용하여 모수의 수를 줄일 수 있다고 설명한다. 또한 학습 할 때, 출력에서 입력으로의 경로 중에서 ResNet의 경우보다 단순한 경로가 존재하므로 기울기 소실 문제를 더욱 효과적으로 해결할 수 있다고 설명한다(Huang *et al.*, 2016).

MobileNet(2017)

MobileNet은 이름에서 유추할 수 있듯이 휴대환경과 내장용 시스템 환경에서 적은 공간적, 시간적 비용으로 동작하도록 개발된 신경망이다. 메모리는 모수의 개수에 비례하므로, 모수의 수를 줄이기 위해 합성곱 계층을 더욱 간소화 시켰다. 다음 [그림 2-10]은 그 방법을 나타낸 것이다.



(a) (b)

[그림 2-10] 일반 합성곱 계층과 인수분해 기법을 이용한 합성곱 계층
(a) 일반 합성곱 계층 (b) 인수분해된 합성곱 계층

입력 채널수가 N, 커널의 크기가 $C\times C$, 출력 채널수가 M이라면 기존의 합성곱 계층에서 모수의 수는 $N\times C^2\times M$이다. MobileNet에서는 채널별로 $C\times C\times 1$ 크기의 커널과 합성곱을 한 후 $1\times 1\times N$ 크기의 커널로 채널간 합성곱을 하여 결과값을 계산하였다. 이 방법을 인수분해(factorization)이라 하며 인수분해의 결과 모수의 수는 $N\times C^2\times M$에서 $C^2\times N+N\times M$개로 감소됨을 알 수 있다. MobileNet은 정확도를 발전시킨 신경망은 아니지만 AlexNet이후의 깊은 합성곱 신경망들이 직면한 많은 모수와 복잡한 구조로 수행시간이 오래 걸리며 메모리가 많이 든다는 문제를 개선한 신경망이다. 이후 MobileNet의 개량형 MobileNet V2를 발표하였다. MobileNet V2는 인수분해 된 합성곱 계층의 커널 사이에 1×1크기의 커널을 이용해 차원을 증가시키고 다시 감소시키는 inverted residual이라는 기법을 사용하여 개량전 신경망보다 높은 정확도를 얻었다고 한다(Sancler, 2018).

NASNet(2017)

Google Brain에서 Auto ML(Auto Machine Learning)이라는 AI로 신경망의 계층 구조 설계와 학습단계에서 사람이 경험적으로 결정해야하는 초모수(hyperparameter)를 자동으로 결정하는 연구를 수행하였다. 그 후 개발된 NASNet을 개발할 때, 신경망을 cell이라는 단위로 나누고 cell의 순서를 바꾸며 조합하는 방식으로 최적화 시켰다. cell은 몇 번째 이전

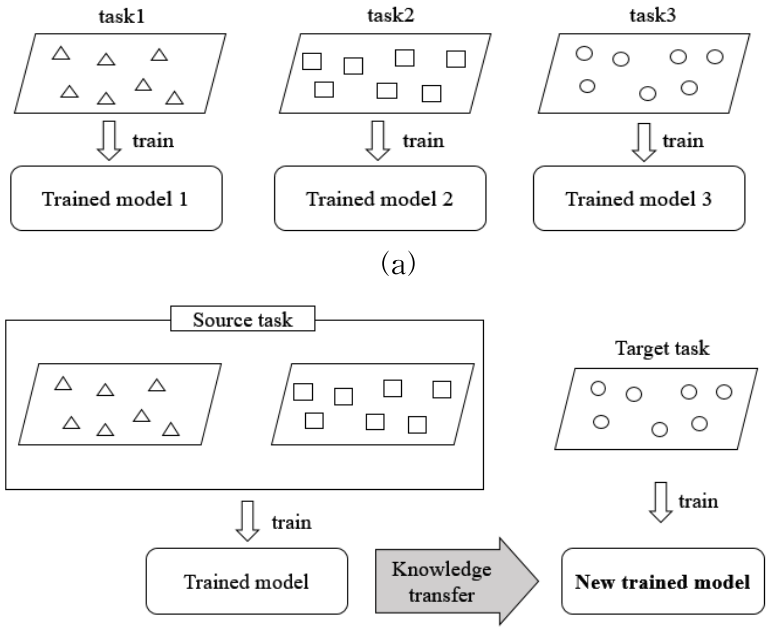
계층을 입력으로 할지, 어떤 연산을 사용할지, 연산 후 값을 더할지 곱할지 등을 조합하여 구성된다. 정확도는 사람이 만든 모델 중 제일 성능이 좋은 신경망과 소수점 두 번째 자리까지 같았다. NASNet은 신경망 구조까지 변경될 수 있는 구조이므로 최적의 모수만 탐색하는 기존의 방식보다 학습에 소모되는 시간이 비교적 크게 증가되었다. NASNet으로 비교적 작은 규모의 영상 데이터베이스인 10개 집단의 5만장으로 구성된 CIFAR-10을 학습하는데 500대의 그래픽 처리 장치를 사용하여 4일이 걸렸다고 한다. NASNet은 cell의 개수를 제한하여 신경망의 크기를 제한할 수 있다. 이때 크기가 크고 정확도가 높은 버전을 NASNet large이라 하고 정확도는 비교적 낮으나 크기가 작은 버전을 NASNet mobile이라 한다(Zoph *et al.*, 2018).

본 논문에서 사용할 심층 합성곱 신경망은 다음 세 가지 기준으로 선택하였다. 1. 가장 단순하며 선행연구에서 다수 사용된 바 있는 초기 모델, 2. 정확도가 70%가 넘으며 시간적, 공간적으로 효율적인 모델, 그리고 3. 시간적, 공간적으로 효율적이지 않더라도 가장 정확한 모델이다. 결과적으로 AlexNet, MobileNet V2 그리고 Inception-ResNet V2를 선택하였다. AlexNet은 실내 위치 추정을 위한 시각적 장소인식에 다수 사용된 바 있으며 가장 기본적인 모델이다. MobileNet V2는 AlexNet보다 15% 정확도 상승을 이룬 약 70%의 정확도를 보이는 모델 중 예측 시간과 모수의 개수가 가장 작기 때문에 선정하였다. Inception-ResNet V2는 전이학습 시 정확도가 가장 높을 것으로 기대되었다. 그 이유를 Kornblith의 연구에서 찾을 수 있다. 해당 연구에 따르면 존재하는 신경망들로 ImageNet의 영상을 분류하거나 다른 대상의 분류를 위해 전이학습 하였을 때, ImageNet의 영상 분류 정확도와 전이학습한 자료의 분류 정확도는 강한 상관관계를 가진다. 하지만 높은 정확도 영역에서 예외적으로 NASNet large 보다 Inception-ResNet V2를 이용하였을 때 평균적으로 전이학습 시 정확도가 더 높게 나왔다(Kornblith *et al.*, 2018).

2.2 전이학습

심층 합성곱 신경망의 모수는 [그림 2-5]를 보면 기본적으로 백만개 이상이므로 다른 기계학습에 비해 대량의 연산량과 학습 데이터가 요구된다. 또한 설계자의 목표에 알맞은 신경망을 설계하기 위해서 계층의 종류와 개수, 초기값 설정, skip connection의 방법 등 고려할 항목이 매우 많고 이들을 가능한 많은 경우로 실험하여 최적화해야하기 때문에 많은 시간과 자원이 필요하다. Auto ML로 인해 신경망 설계조차 컴퓨터로 자동화 할 수 있으나, 최적화할 항목이 더욱 증가하여 간단한 데이터에 대해 학습할 때도 더욱 많은 자원이 소모되므로 범용성이 떨어진다. 결과적으로 위와 같이 심층 합성곱 신경망의 구조를 설계하고 설계된 신경망을 초기단계부터 학습시키는 것은 많은 시간과 자원이 요구된다는 문제가 존재한다. 이를 해결하기 위한 방법으로 특정 대상을 위해 구현된 알고리즘을 다른 비슷한 분야의 대상에 응용하는 방식인 전이학습이 높은 효율성과 우수한 성능이 입증되어 많은 분야에서 사용되고 있다.

전이학습은 신경망뿐만 아니라 일반적인 기계학습에서 사용되어 온 기법이다. [그림 2-11]은 전이학습을 사용하지 않은 기계학습과 전이학습을 사용한 기계학습의 차이점을 표현한 그림이다.



[그림 2-11] 일반 기계학습과 전이학습을 이용한 기계학습의 차이
(a) 일반 기계학습 (b) 전이학습을 이용한 기계학습

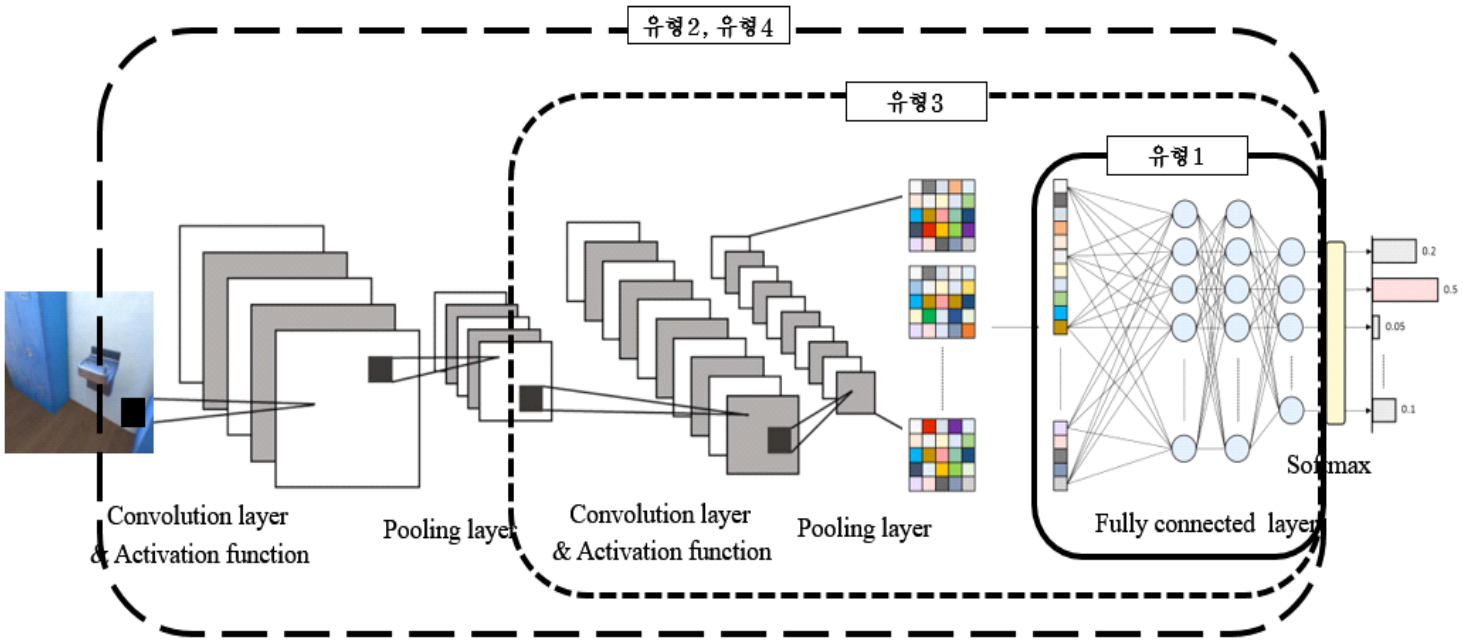
[그림 2-11] (a)와 같이 일반적인 기계학습 방법을 이용하면, 주어진 데이터에 대해서 각각의 모델이 독립적으로 구현된다. 반면에 [그림 2-11] (b)와 같이 전이학습 기반의 학습 방법을 사용할 경우는 대상 데이터와 비슷하지만 다른 유형의 학습 데이터를 이용하여 모델의 선행학습을 수행한다. 그 후 지식 이전(knowledge transfer)을 통하여 선행학습된 모델의 구조와 모수를 그대로 가져오고 대상 데이터를 가져온 모델에 학습시켜 새로운 모델을 구현하는 방식이 적용된다.

전이학습 기반의 심층 신경망 학습 방법은 대상 데이터의 보유량과 선행 학습에 사용된 데이터와의 유사도를 기준으로 [표 2-1]에 표현된 네 가지 유형이 존재한다(이한수 *et al.*, 2018).

[표 2-1] 대상 데이터와 선행 학습 데이터의 유사도와
대상 데이터 수량에 따른 전이학습의 유형 분류

데이터 유사도 / 데이터 수량	높음	낮음
적음	유형 1	유형 3
많음	유형 2	유형 4

신경망에서 전이학습은 대량의 데이터로 선행 학습된 신경망의 구조와 모수를 불러온 후 유형별로 학습시킬 모수의 범위를 설정하여 미세 조정하는 방식으로 이루어진다. 미세 조정이란 작은 학습률을 설정하여 모수를 수정한다는 의미로, 전이학습 시, 선행 학습된 모수들은 최적 모수와 차이가 적다는 가정하에 작은 학습률을 설정한다. 학습률은 오류 역전파 알고리즘을 이용하여 모수를 수정할 때 수정치에 곱해주는 값으로 0에 근접할수록 모수의 수정값이 작아지며, 0.001이하의 값이 미세 조정에 사용된다(Zoph *et al*, 2018). [그림 2-12]는 합성곱 신경망을 지식 이전한 후, 유형별 학습 단계에서 수정되는 모수의 범위이다.



[그림 2-12] 전이학습 유형별 수정되는 모수의 범위

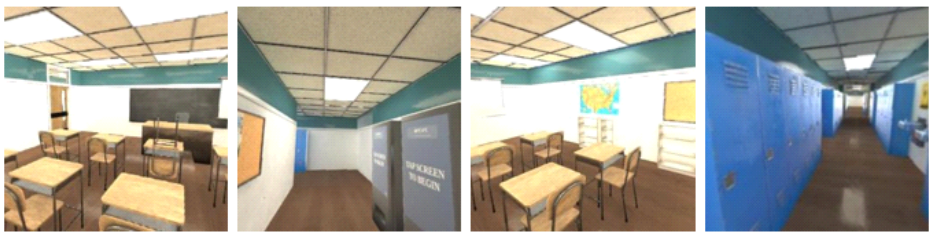
유형1은 대상 데이터의 보유량이 적기 때문에 선행 학습된 심층 신경망을 전체적으로 미세 조정할 경우에는 과적합이 발생할 수 있다. 따라서 전역 연결 계층이전의 특징 추출을 위한 계층은 학습하지 않고 전역 연결 계층의 모수만 학습하여 최종 모델을 구현한다. 이는 대상 데이터와 선행 학습에 사용된 데이터의 유사도가 높으므로 같은 특징을 추출하고 분류기만 학습시켜 모델을 구현한다는 의미이다.

유형2와 4의 경우에는 대상 데이터의 보유량이 많기 때문에 과적합의 위험이 적으므로 전체 신경망에 대해 미세조정을 실시하여 최종 모델을 구현한다. 이는 대상 데이터의 분류를 위해 모든 계층들이 수정되고, 그 계층들에서 최적화된 특징들을 추출한다는 의미이다.

유형3은 전역 연결 계층과 이에 가까운 계층의 모수를 수정하고 입력과 가까운 계층의 모수는 수정하지 않는다. 즉, 낮은 단계의 특징은 선행 학습된 모델의 것을 사용하고, 이는 데이터 유형이 다소 다르지만 낮은 단계에서 추출되는 특징(점, 경계, 색상차이 등)은 선행 데이터와 대상 데이터의 분류에 공통적으로 사용한다는 의미이다.

선행 학습 모델로 ImageNet의 영상들을 이용하여 학습된 심층 신경망을 전이학습 할 때, 유형1~4중 하나를 선택하여야 한다. 데이터 간 유사도가 높고 낮음과 데이터의 수량이 많고 적음이라는 기준은 절대적인 수치가 없으므로 유형을 선정하는 것은 시스템 구축자의 재량에 달려있다. 데이터 간 유사도는 각각의 영상의 어떤 센서를 통해 얻어졌는지, 영상의 구도와 시점, 영상안의 물체 개수 등을 종합하여 판단된다. 다음은 ImageNet의 예시 영상과 본 연구의 시각적 장소인식에 이용한 영상의 예시이다.

(a) ImageNet의 영상



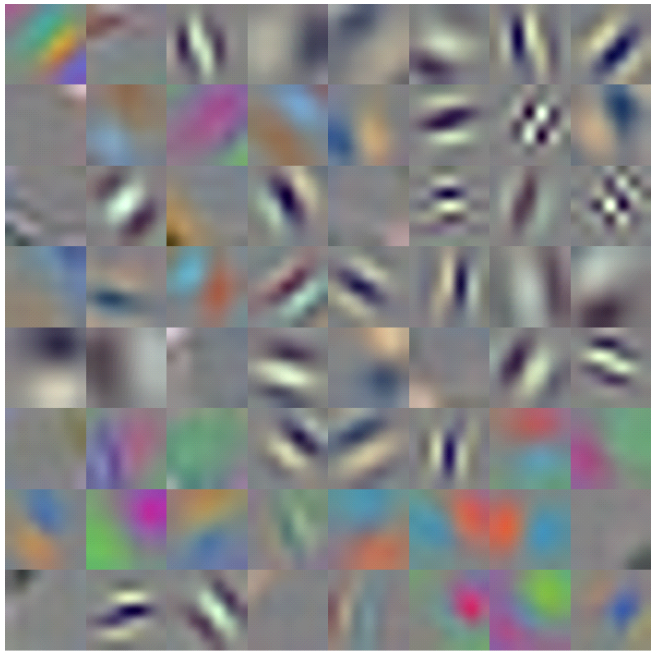
(b) 본 실험에 사용할 실내 장소 영상

[그림 2-13] 학습에 이용된 영상들

본 연구에서 선행 모델로 사용할 합성곱 신경망들은 검색엔진에서 수집된 영상들로 구축한 ImageNet 데이터베이스로 학습된다. [그림 2-13] (a)는 ImageNet의 예시 영상들을 보여준다. 이들은 사물들을 여러 구도와 환경에서 촬영한 RGB영상이다. 그리고 시각적 장소인식을 이용한 실내 위치 추정에서 사용되는 영상은 [그림 2-13] (b)와 같이 실내 환경에 사물이 존재하는 RGB 영상이다. 선행연구 중에는 이 두 영상이 RGB 영상이므로 유사도가 있다고 판단하여 전역 연결 계층 이전의 계층을 그대로 사용하는 유형1의 방법을 사용한 연구가 있다(Werner *et al.*, 2016, Kim *et al.,* 2015). 반면에 둘 다 RGB 영상이더라도 촬영한 목적이 다르고 구도가 다르므로 유사도가 크지 않다고 판단하여 전역 연결 계층 이전의 계층도 학습하는 유형3의 방법을 사용한 연구도 존재한다(Zhang *et al.*, 2016). 따라서 본 연구에서는 어떤 유형의 전이학습에서 실내 위치 추정이 더 정확한지 알아보기 위해 신경망 모델별로 유형1과 유형3의 방법을 모두 사용하여 분석한다.

유형3의 경우는 학습시키지 않을 계층의 범위를 결정하는 과정이 필요

하다. Szegedy는 GoogLeNet의 7×7 크기의 초반 3개의 합성곱 계층은 영상의 점, 경계, 색상 등 낮은 단계의 특징 추출에 사용되어 다른 버전의 GoogLeNet에서도 건드리지 않았다고 한다(Szegedy *et al.*, 2016). 다음은 GoogLeNet의 처음 필터 64개를 시각화 한 그림이다. 이 필터들은 영상에서 방향의 경계, 색상차이 등을 추출하는 필터임을 알 수 있다.

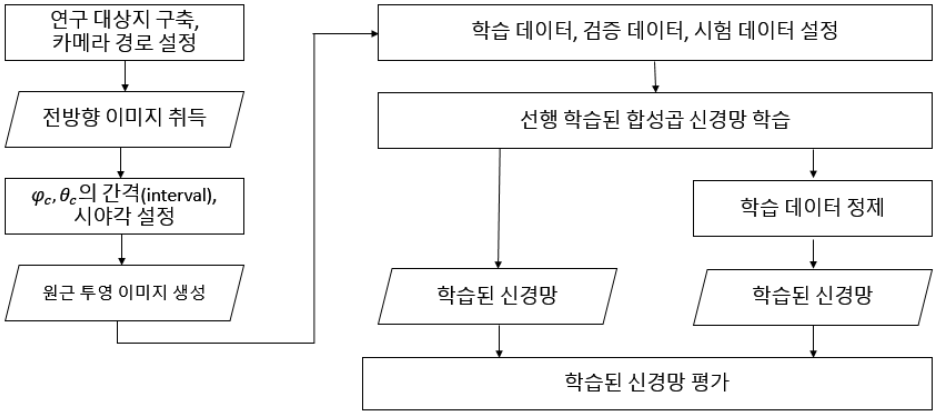


[그림 2-14] GoogLeNet의 첫 합성곱 계층의 필터 64개

그러므로 전이학습 유형3를 기반으로 학습할 때, AlexNet의 입력에서 가장 가까운 첫 번째 합성곱 계층은 기본정보를 추출하는 계층이고 점, 경계, 색 등의 기본 정보는 실내사진의 특징 추출에도 사용된다고 판단하여 학습시키지 않았다. 단, MobileNet V2와 Inception-ResNet V2는 기본 정보를 추출할 때에도 AlexNet처럼 한 개의 필터가 아니라 3×3이나 3×1의 작은 필터를 사용하기 때문에 이를 반영하여 각각 입력에서 가장 가까운 3개, 5개의 합성곱 계층을 학습시키지 않았다.

3. 실험 방법

3장에서는 실험 방법을 설명한다. 다음 [그림 3-1]의 순서도와 같이 연구 대상지 구축, 카메라 경로 설정, 원근 투영 영상 이미지 생성, 합성곱 신경망 학습, 학습 데이터 정제, 그리고 평가방법을 설명하는 내용으로 구성되어있다.

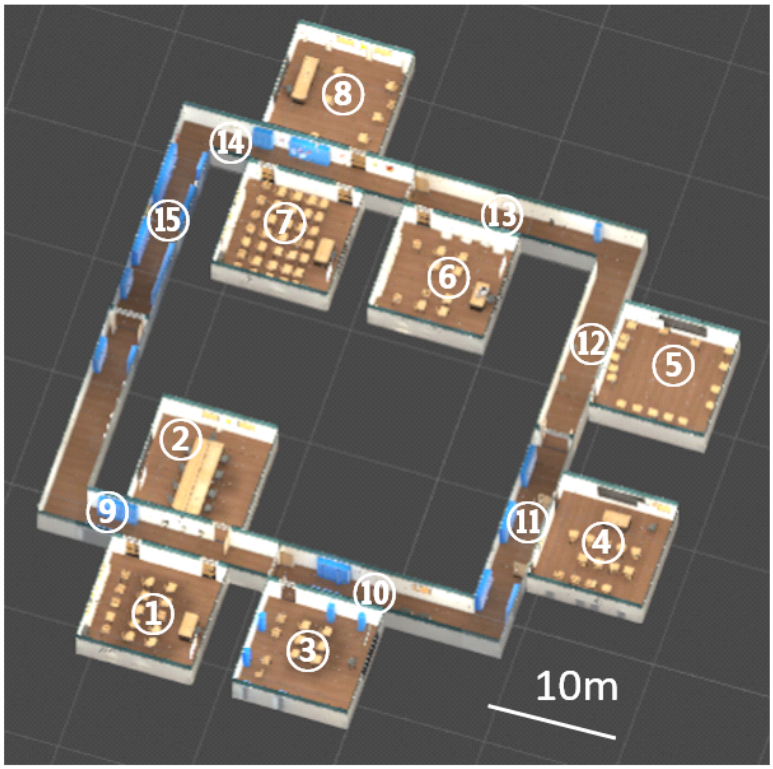


[그림 3-1] 실험 순서도

먼저 실험은 가상환경에서 진행되었기 때문에 대상지를 구축하는데 사용한 프로그램, 구축할 때 고려한 점, 카메라의 경로를 설명한다. 그 다음, 영상 데이터베이스 구축에서는 실험에서 사용한 전방향 카메라와 얻어지는 전방향 영상, 그리고 전방향 영상을 원근 투영 영상으로 변환하는 방법을 다룬다. 이 영상들을 이용하여 2장에서 선택한 사전학습 된 합성곱 신경망들을 전이학습 유형별로 학습시킬 때, 원근 투영 영상들을 학습, 검증, 그리고 시험 데이터로 나눈 방법과 설정한 초모수(hyper parameter) 값들을 살펴본다. 그 후, 학습 데이터 정제에서는 합성곱 신경망 학습 영상, 검증 영상에서 발견된 정보가 부족한 영상을 제거하는 방법을 제안한다. 마지막으로 분석방법에서는 합성곱 신경망의 평가를 위한 기준이 포함된 분석방법을 설명한다.

3.1 실험 대상지 구축

실험은 게임 개발 도구 프로그램인 Unity를 이용하여 구축된 가상환경에서 실행되었다. Unity를 이용하면 가상환경에서 물체의 재질에 따른 반사정보와 빛 퍼짐을 세밀하게 조정하고 3D 객체를 만들고 배치할 수 있다. 또한 다른 사용자가 제작한 3D모형들을 불러오고 개체를 재배열하여 원하는 환경을 쉽게 제작할 수 있다. 추가로 카메라의 경로, 종류, 촬영 방법 등을 다양하게 설정할 수 있다.



[그림 3-2] 구축한 가상 환경의 전경

분류하고자 하는 구역의 개수는 15개로 설정하였다. 각 구역의 넓이는 $100m^2$으로 모두 동일하다. [그림 3-2]는 가상공간 전체와 구역을 표시한 그림이다. 구역 1부터 구역 8은 교실이며 구역 9부터 구역 15는 복도로 구성하였다. 현실의 장소들은 장소들 간 공통점과 각 구역의 특징을 모두 가지고 있다(Lowry *et al.*, 2015). 이 사항을 반영하여, 벽과 바닥은 교실 구역들과 복도 구역들이 같으며 [그림 3-3]에 존재하는 가구들은 구역 2를 제외한 교실들에 존재한다는 공통점을 설정하였다.

[그림 3-3] 공통으로 존재하는 의자, 책상, 칠판

또한 몇몇 구역에는 사물함과 음수대가 존재하는데, 사물함과 음수대는 같은 종류를 배치하여 장소 간 공통점으로 설정하였다. 다음 [그림 3-4]는 같은 종류로 설치된 사물함과 음수대의 모습이다.

[그림 3-4] 공통으로 존재하는 사물함, 음수대

추가적으로 각 구역을 구분지어 주는 특징을 [그림 3-5]와 같이 부여하였다.

(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

[그림 3-5] 각 장소의 특징들
(a) 구역 2 (b) 구역 3 (c) 구역 5 (d) 구역 7 (e) 구역 10
(f) 구역 12 (g) 구역 13 (h) 구역 14 (i) 구역 15

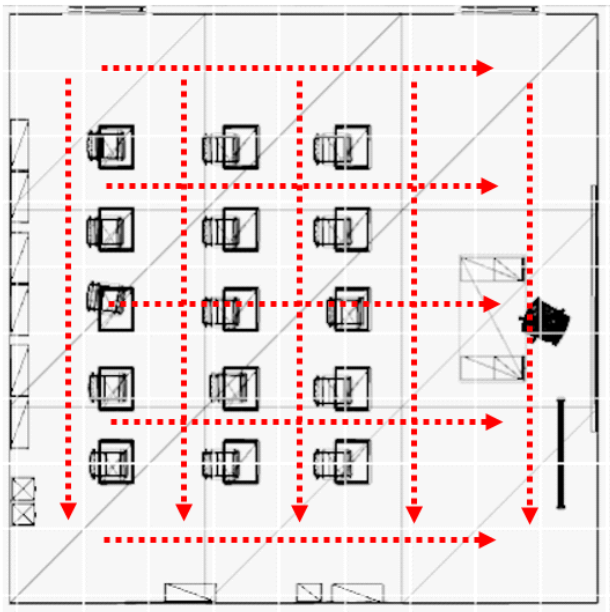
[그림 3-5] (a)의 구역 2는 크기가 큰 회의용 책상이 교실 중앙에 존재한다. [그림 3-5] (b)의 구역 3은 다른 교실에 없는 사물함이 존재한다. [그림 3-5] (c)의 구역 5는 비어있는 공간이 많은 교실이다. [그림 3-5] (d)인 구역 7은 가구가 다른 곳에 비해 많다. [그림 3-5] (e)는 구

역 10이며 키오스크가 있다. [그림 3-5] (f),(g)는 각각 구역 12, 구역 13이며 비어있는 공간이 많은 복도이다. [그림 3-5] (h)는 구역 14이며 포스터가 사물함에 붙어있다. [그림 3-5] (i)는 구역 15로 많은 사물함이 복도의 대부분을 차지한다.

3.2 카메라 경로 설정

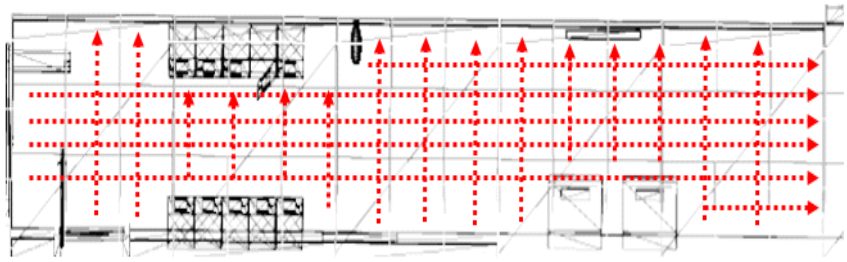
선행 연구에서 실제세계의 실내 위치 추정을 위한 합성곱 신경망의 학습 영상들은 동영상을 촬영한 후, 동영상들을 이루는 프레임들로 생성된 바 있다. 가상 환경에서도 같은 방법으로 카메라 경로를 지정 후, 동영상 촬영 방식과 동일하게 초당 촬영할 프레임수를 설정하여 카메라가 경로를 이동하며 연속적으로 영상을 촬영하도록 설정하였다. 또한 구역별 학습 데이터의 불균형으로 인한 효과를 고려하지 않기 위해 모든 구역의 학습 데이터 수는 같도록 설정하였다.

구역이 교실일 경우(구역 1부터 구역 8), 교실의 모든 부분을 고르게 촬영하기 위해, [그림 3-6]과 같이 격자 형태로 이동경로를 지정하였다. 단, 경로상에 물체가 있는 경우, 물체에 닿지 않게 높이 값을 수정하였다.



[그림 3-6] 교실 구역에서의 카메라 경로

복도 구역 또한 모든 부분을 고르게 촬영하도록 [그림 3-7]과 같이 카메라 경로를 설정 하였다. 단, 장애물이 많아 전체 경로가 짧은 경우는 전체적으로 이동속도를 낮추어 구역별 학습 영상 개수가 모두 동일하도록 조치하였다.



[그림 3-7] 복도 구역에서의 카메라 경로

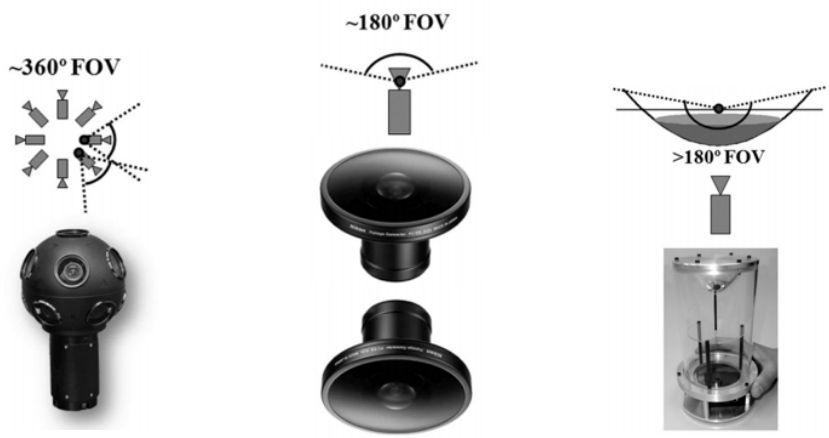
실험 환경이 가상 환경이라도 사람이 촬영하는 것으로 가정하여 카메라가 3축으로 -5°에서 5° 만큼 무작위로 회전하도록 설정하였고, 높이 또한 1m에서 2m 사이에서 변화하도록 설정하였다.

3.3 영상 데이터베이스 구축

영상 데이터베이스 구축은 전방향 영상을 취득하고 이를 원근 투영 영상으로 분할하는 과정으로 이루어져있다.

3.3.1 전방향 영상 취득

전방향 영상은 360°의 수평 시야각(FOV(Field of View)), 혹은 완전 구형의 시야가 담긴 영상이다(Scaramuzza, 2014). 전방향 영상을 획득하는 방법은 [그림 3-8] (a)처럼 일반 카메라 영상을 여러 장 결합하는 방법, [그림 3-8] (b)의 반구(Hemishpere) 이상인 넓은 시각의 렌즈를 사용하는 방법, 그리고 [그림 3-8] (c)의 거울반사형 시각센서(Catadioptric Vision Sensor)를 이용하는 방법이 있다(양윤모, 2004).

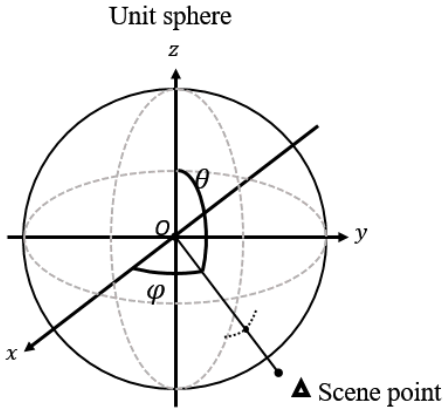


(a) (b) (c)

[그림 3-8] 전방향 카메라의 종류
(a) 일반 카메라 영상 여러 장을 결합
(b) 넓은 시각의 렌즈 사용
(c)거울반사형 시각 센서 사용
(Scaramuzza, 2014)

핀홀 카메라(pinhole camera) 모형에서 얻어지는 원근 투영 영상에 비해 전방향 영상은 한 장에 더 많은 주변 정보를 얻을 수 있기 때문에 감시 및 로봇 분야 등에서 다양하게 사용된다(Scaramuzza, 2014). 많은 정보가 포함된 전방향 영상 한 장을 분할하면 원근 투영 영상 다수를 얻을 수 있고, 이 방법으로 대량의 학습데이터가 요구된다는 심층 합성곱 신경망의 단점을 극복할 수 있다.

모든 수직, 수평 구간이 촬영된 전방향 영상은 [그림 3-9]와 같이 주변의 색상정보가 단위 구면에 대응되도록 모형화 할 수 있고, 이 단위 구는 구면좌표계로 표현할 수 있다(Aghayari *et al.*, 2017).

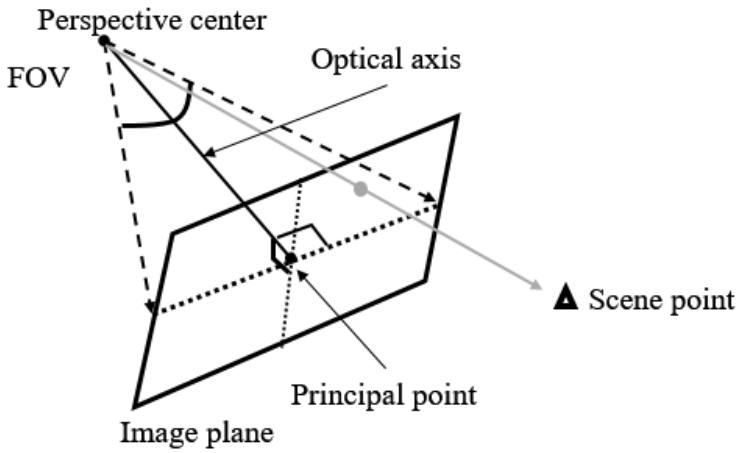


[그림 3-9] 구면에 색상정보가 대응된 전방향 영상 모형

[그림 3-9]에서 표현된 주변 한 점(Scene point)의 색상정보는 구면의 중심을 잇는 직선과 단위 구면이 만나는 점에 대응되고, 이 점은 양의 x축과 이루는 수평각 $\varphi(0 \le \varphi < 2\pi)$, 양의 z축과 이루는 수직각 θ $(0 \le \theta \le \pi)$, 즉 (φ, θ)로 표현할 수 있다. 그러므로 전방향 영상 한 장에는 단위 구면 모형으로 주변의 모든 색상정보가 (φ, θ)로 대응되어 있다.

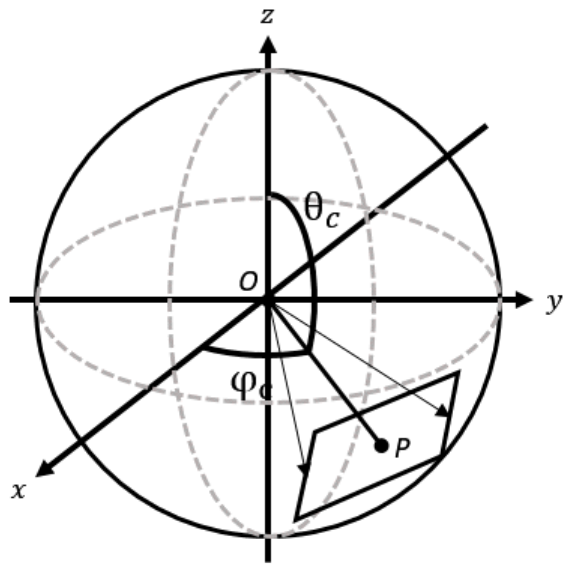
3.3.2 원근 투영 영상 취득

[그림 3-10]은 일반 원근 투영 영상을 얻는 핀홀 카메라의 기하학적 모형을 나타낸 그림이다.

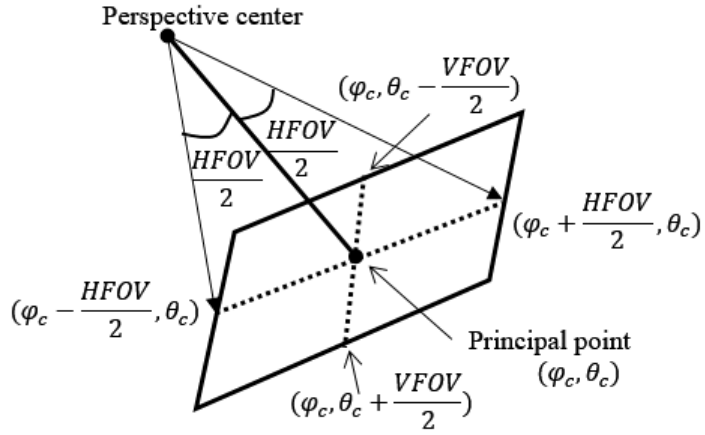


[그림 3-10] 핀홀 카메라의 기하학적 모델

[그림 3-10]은 원근 투영 중심점(perspective center), 영상 평면(image plane), 주점(principal point), 그리고 광축(optical axis)을 표시한 핀홀 카메라 모형이다. 이 모형에서 투영 중심점, 광축, 투영 중심점과 영상 평면의 거리인 초점 거리, 수평 시야각(HFOV), 그리고 수직 시야각(VFOV)을 설정하면 영상 평면의 크기가 정해져 핀홀 카메라의 모형을 생성할 수 있다. 또한 투영 중심점, 영상에서 물체의 한 점, 물체의 한 점이 한 직선 위에 존재한다는 공선 조건을 이용하면 주변 환경의 색상 정보를 가진 영상을 생성할 수 있다. 위의 모형을 이용하여 [그림 3-11]과 같이 구의 형태로 매핑 된 전방향 영상에서 여러 장의 원근 투영 영상을 생성할 수 있다.



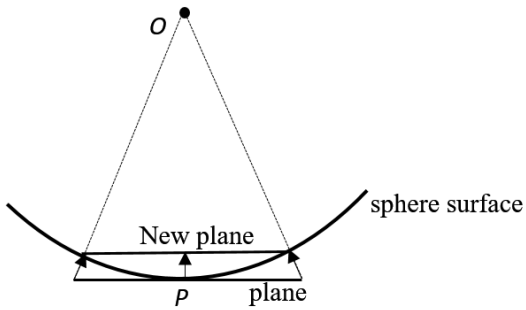
(a)



(b)

[그림 3-11] 구면에 색상정보가 매핑 된 전방향 영상에서
원근 투영 모델
(a) 구면 전방향 영상에서의 원근 투영 모델
(b) 원근 투영 모델 확대

전방향 영상 모형의 단위 구에서 각도(φ_c, θ_c)를 설정하고 이에 대응하는 구면상의 점P를 설정한다. [그림 3-11]의 (b)는 해당 부분을 확대한 것이다. 이때 P에서 단위 구에 접하는 평면을 생각할 수 있다. 설정된 수평 시야각과 수직 시야각에 따라 단위 구에 접하는 평면의 크기를 제한할 수 있다. 이 상황에서 구의 중심점O를 투영 중심점, 구면상의 점P를 주점, $\overline{OP}$와 그 거리를 광축과 초점거리, 그리고 P에서 단위 구에 접하며 크기가 제한된 평면을 영상 평면으로 볼 수 있으므로 핀홀 카메라 모형을 구성 할 수 있다. [그림 3-12]와 같이 영상 평면의 크기를 조정하고 구의 중심으로 평행이동을 수행하면 영상 평면의 네 꼭짓점이 구면상에 위치 할 수 있으며, 규모만 작아질 뿐, 핀홀 카메라의 요소들은 그대로 존재한다.



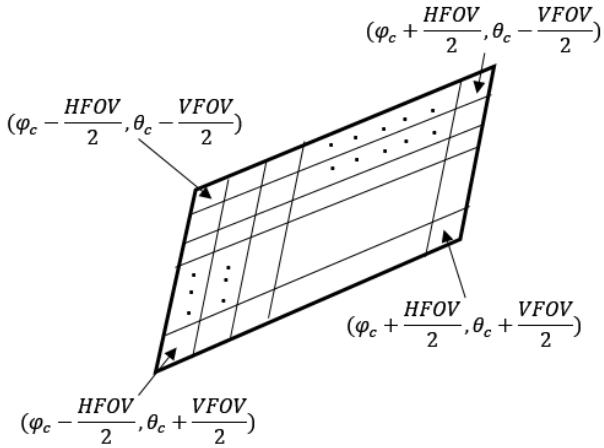
[그림 3-12] 영상 평면을 단위 구 내부에 위치하도록
규모 축소 및 이동

다음은 단위 구에서 구성된 핀홀 카메라 모형의 영상 평면에 색상을 대응하는 방법이다. 이는 영상 평면 격자화 및 각 픽셀에 좌표를 부여하는 단계, 그리고 구면에서 색상 정보를 리샘플링(resampling)하는 단계, 이렇게 두 단계로 이루어진다.

먼저 영상 평면 격자화 및 각 픽셀에 좌표를 부여하는 단계의 과정을 설명한다. 평면의 각 꼭지점에서 [그림 3-11] (b)을 이용하면 각각 구면 좌표의 각도값은 [그림 3-13]과 같이

$(\varphi_c - \frac{HFOV}{2}, \theta_c - \frac{VFOV}{2})$, $(\varphi_c + \frac{HFOV}{2}, \theta_c - \frac{VFOV}{2})$

,$(\varphi_c - \frac{HFOV}{2}, \theta_c + \frac{VFOV}{2})$, $(\varphi_c + \frac{HFOV}{2}, \theta_c + \frac{VFOV}{2})$임을 알 수 있다.



[그림 3-13] 평면 격자화 및 네 꼭지점의 좌표 부여

영상 평면의 픽셀들의 좌표들은 일정한 x,y,z값의 배수만큼 차이가 있다. 그러므로 구면 좌표값들을 직교 좌표계의 값으로 변환하는 과정을 거친다. 먼저 네 꼭지점의 각도 φ,θ와 거리 r을 x,y,z의 값으로 변환하는 관계식은 수식(3-1)과 같다. 이 점들은 [그림 3-12]에 표현된 과정으로 단위 구면상에 있으므로 모두 O과의 거리 r은 1이다.

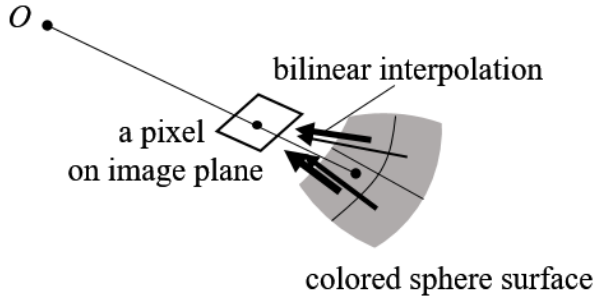
$$\begin{aligned} x &= r\sin\theta\cos\varphi \\ y &= r\sin\theta\sin e\varphi \\ z &= r\cos\theta \end{aligned} \qquad (3\text{-}1)$$

네 꼭지점을 직교좌표로 변환하고 평면 영상을 이루는 가로 세로의 픽셀수를 결정하면 이웃하는 픽셀간의 x,y,z값의 차이를 알 수 있고 평면 영상에 존재하는 모든 내부 픽셀들의 좌표를 구할 수 있다.

다음 단계는 구면에서 영상 평면으로 색상 정보를 리샘플링하는 단계이다. 색상정보는 전방향 영상의 구면상에 존재하고 구면상의 각도값 (φ,θ)으로 대응되어 있으므로 평면상의 직교 좌표값을 다시 구면 좌표값으로 변환한다. 수식(3-2)는 직교좌표계의 좌표(x,y,z)를 구면좌표계의 좌표(r,φ,θ)로 변환하는 관계식이다.

$$\begin{aligned} r &= \sqrt{x^2+y^2+z^2} \\ \varphi &= \arctan\frac{y}{x} \\ \theta &= \arctan\frac{z}{r} \end{aligned} \qquad (3\text{-}2)$$

식(3-2)로 얻은 각도에 해당하는 구면상의 색상정보를 취득한다. 단 구면은 이산적인 픽셀들로 이루어져 있으므로 구면상의 색상을 취득할 때 [그림 3-14]와 같이 주변의 가장 가까운 네 개 픽셀의 색상 정보를 이용하여 쌍선형 보간법(bilinear interpolation)으로 취득한다.



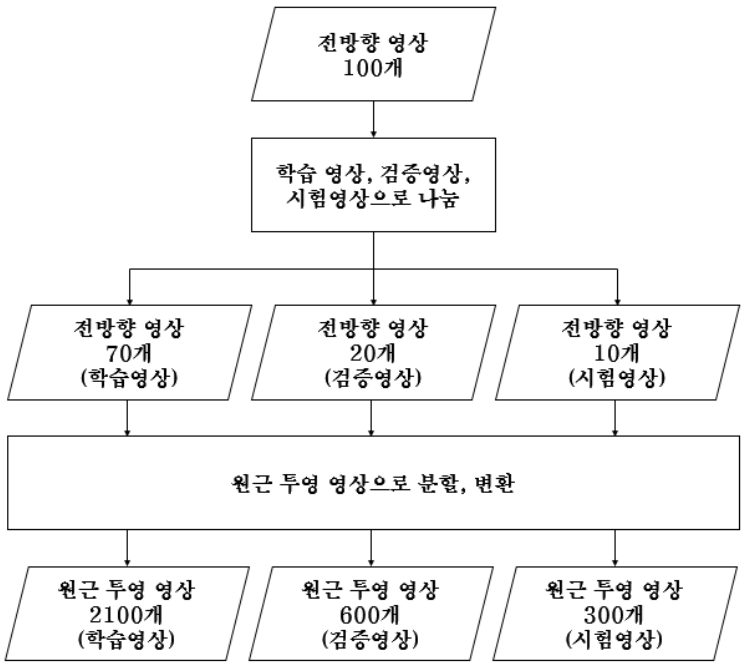
[그림 3-14] 구면에서 영상 평면으로 색상정보 추출

종합하면 전방향 영상에 주변 색상정보가 매핑된 단위 구가 있고 수직, 수평 시야각, 영상 평면의 가로 그리고 세로 픽셀 수를 설정하면 광축의 각도(φ_c,θ_c)가 주어질 때마다 주변 환경의 색상정보 일부를 포함한 원근 투영 영상 한 장이 생성된다. 한 장의 전방향 영상에서 광축의 각도(φ_c,θ_c)는 다수 설정할 수 있고 그 만큼 원근 투영 영상들이 생성된다. 이 영상들은 심층 합성곱 신경망 학습에 사용할 수 있으므로, 언급한 일련의 원근 투영 영상 생성 과정은 다수의 학습영상을 취득하는 방법으로 사용될 수 있다. 또한 부수적으로 모든 수평 시야를 가진 전방향 영상을 사용하였으므로 주변의 시야를 고르게 촬영한 효과도 가진다.

본 연구에서는 현재 사용되는 스마트폰 카메라의 시야각을 고려하고, 합성곱 신경망은 정사각형의 이미지를 입력으로 받기 때문에 원근 투영 영상의 수직, 수평 시야각을 70°로 설정하였다. 또한 원근 투영 카메라 광축의 수평각인 φ의 간격은 36°로 설정하여 종중복도가 48.5%인 영상들이 수평으로 1회전 마다 10장씩 생성된다. 광축의 수직각 θ는 바닥과 천장에 구역을 식별할 만한 정보가 없으므로 60°, 90°, 120°로만 설정하여 횡중복도가 51.1%인 영상들이 되도록 설정하였다. 그러므로 한 장의 전방향 영상에서 10×3 = 30장의 원근 투영 영상이 생성된다. 본 연구의 실험에서 전방향 영상을 구역 별로 100장씩 촬영하였다. 결과적으로 구역별 100장의 전방향 영상에서 3,000장의 원근 투영 영상이 취득되었으며 모든 구역에서 45,000장의 원근 투영 영상을 생성하였다.

3.4 전이학습을 이용한 심층 신경망 학습

2절에서 언급하였듯 ImageNet의 영상으로 사전 학습된 AlexNet, MobileNet V2, 그리고 Inception-Resnet V2의 구조와 모수를 지식 이전하여 실내 위치 추정에 이용할 합성곱 신경망을 학습시켰다. 학습에 사용할 데이터로 생성된 원근 투영 영상을 사용하였다. 이때, 촬영된 구역당 100개의 취득된 전방향 영상을 [그림 3-15]와 같이 학습 영상, 검증 영상, 그리고 시험 영상으로 나누기 위해 70장, 20장, 10장씩 분리하였다. 그 후, 각 영상들을 원근 투영 영상 2,100장, 600장, 그리고 300장으로 분할하였다.



[그림 3-15] 취득한 영상을 학습, 검증, 시험영상으로 분리하는 방법

훈련, 검증 그리고 시험 영상을 원근 투영 영상 만들기 전단계인 전방향 영상에서 분리한 이유는 각각 다른 위치에서 촬영된 효과를 얻기 위함이다. 각 영상들이 각 다른 위치에서 촬영되는 것이 실제세계의 실내

위치 추정에서 알맞은 상황이다. 만약 같은 위치에서 촬영된 영상들이 학습 영상과 시험 영상에 동시에 포함된다면 횡중복도와 종중복도가 약 50%이기 때문에 학습 영상에 시험 영상이 들어가는 효과를 나타내어 정확도가 과대평가 될 수 있다.

신경망을 학습 시킬 때 사람이 정해주어야 하는 초모수는 학습률(learning rate), 최대 반복 횟수(max epoch), 한 번에 학습 시킬 이미지의 개수(mini batch size)가 있다.

먼저 학습률의 경우, 지식 이전된 합성신경망을 미세조정 할 때 이미 많은 모수들이 학습되어 있으므로 0.001이하의 낮은 학습률을 설정한다(Zoph *et al*, 2018). 반복 실험 결과, 본 실험에서는 학습률에 따른 큰 차이가 발생하지 않아 프로그램의 기본값인 0.0001을 사용하였다. 단, 출력의 개수가 달라짐에 따라 전역 연결 계층의 모수들은 초기상태부터 학습되기 때문에 전역 연결 계층의 학습률만 설정된 학습률의 10배 값으로 설정하였다. 또한 학습이 진행되는 과정에서 모수가 최적값으로 수렴됨에 따라, 미세한 범위내에서 수정되어야 하므로 학습률을 감소시킬 필요가 있다. 그러므로 한 번의 반복 후 학습률에 0.01을 곱하여 학습률을 감소시키는 기법을 사용하였다.

학습 영상 전체를 1회 학습하였을 때 1반복(1 epoch)이 종료된 것으로 본다. 이렇게 학습 영상 전체를 몇 회에 걸쳐 학습할 것 인지 설정하는 초모수가 최대 반복 횟수이다. 신경망의 학습에서는 학습 정확도와 검증 정확도가 수렴하였다고 판단할 때까지 최대 반복 횟수를 증가시켜야 한다. 전이학습의 경우, 신경망을 처음부터 학습시키는 경우 보다 정확도가 빠르게 수렴한다. 본 연구의 학습에서 정확도는 2회 반복 학습 후부터 수렴하는 경향을 보였기 때문에, 최대 반복 횟수는 3회로 설정하였다. 단, 몇몇 경우에 검증 영상의 손실이 증가하는 과적합의 경향을 보여, 해당 경우에 2회만 학습하는 조기 종료기법을 사용하였다.

신경망의 학습은 모수들의 수정량을 학습 데이터 한 개 마다 반영하는 것이 아니라 일정한 개수의 데이터들의 수정량을 더하여 한 번에 수정한다. 이때 '일정한 개수'가 mini batch size이다. 연구에 의하면 mini

batch size는 30 부근의 값에서 학습한 신경망이 평균적으로 좋은 성능을 보였다(Goyal *et al.*, 2018). 그러므로 프로그램의 기본값인 30을 사용하였다.

2.2.2 전이학습에서 언급하였듯, 신경망마다 유형1과 유형3 상황을 실험하였다. 유형3의 경우 AlexNet의 경우 첫 번째 합성곱 계층, MobileNet V2와 Inception-ResNet V2는 각각 세 번째와 다섯 번째 합성곱 계층까지 학습률을 0으로 설정하였다. 정리하자면 실험에 사용된 초모수는 [표 3-1]과 같다.

추가적으로 본 연구의 실험에 사용된 컴퓨터의 사양은 Intel i7-6700k CPU, 32GB RAM, Nvidia GTX 1080ti이며, 프로그램은 Matlab 2019a의 deep learning toolbox를 사용하였다.

[표 3-1] 학습 시 설정한 초모수 값

적용한 전이학습 / 적용한 합성곱 신경망 모형	유형1	유형3
AlexNet	학습률 : 0.0001 학습률 감소 계수 : 0.01 최대반복 : 3 mini batch size : 30 학습 범위 : 전역 연결 계층만 학습	학습률 : 0.0001 학습률 감소 계수 : 0.01 최대반복 : 3 mini batch size : 30 학습 범위 : 첫 번째 합성곱 계층을 제외한 계층들을 학습
MobileNet V2	학습률 : 0.0001 학습률 감소 계수 : 0.01 최대반복 : 3 mini batch size : 30 학습 범위 : 전역 연결 계층만 학습	학습률 : 0.0001 학습률 감소 계수 : 0.01 최대반복 : 2 mini batch size : 30 학습 범위 : 세 번째 합성곱 계층까지 제외한 계층들을 학습
Inception-ResNet V2	학습률 : 0.0001 학습률 감소 계수 : 0.01 최대반복 : 3 mini batch size : 30 학습 범위 : 전역 연결 계층만 학습	학습률 : 0.0001 학습률 감소 계수 : 0.01 최대반복 : 2 mini batch size : 30 학습 범위 : 다섯 번째 합성곱 계층까지 제외한 계층들을 학습

3.5 데이터 정제

3.5 데이터 정제에서는 정보가 적은 데이터를 제거하는 방법을 다룬다. 먼저 데이터 정제를 고려한 동기에 대해 설명한다. 그리고 엔트로피와 정제하지 않은 영상 데이터베이스로 학습시킨 신경망의 분류결과 확률을 이용한 데이터 정제 방법을 제안한다.

3.5.1 동기

심층 신경망의 학습뿐만 아니라 일반적인 기계학습에서 학습 데이터의 품질은 모델에 큰 영향을 끼친다. 그러므로 모델 학습의 전처리에서 잡음이 심한 데이터, 경향에 맞지 않는 이상치 데이터, 혹은 잘못 분류 된 데이터를 제거하거나 수정하는 작업은 반드시 필요하다. 본 연구에서 실험은 가상 환경에서 실행되었기 때문에 움직임에 의해 흐려진 영상, 라벨링이 잘못된 데이터, 잡음이 심한 데이터는 없으므로 이들에 대한 학습 데이터 정제는 필요하지 않다.

하지만 학습 영상, 검증 영상 및 시험 영상으로 사용할 영상 데이터를 관찰하던 중 [그림 3-16] 처럼 '정보가 부족한' 영상들이 다수 관측되었다.

[그림 3-16] 적은 정보로 시각적 실내 위치 추정이 불가능한 영상

이러한 구역별 공통으로 존재하는 빈 벽 혹은 칠판만을 포함하고 다른 대상이나 특징점이 없는 영상은 신경망이 극단적으로 과적합이 되지 않는 한, 시각적 방법으로 분류가 불가능한 영상이다.

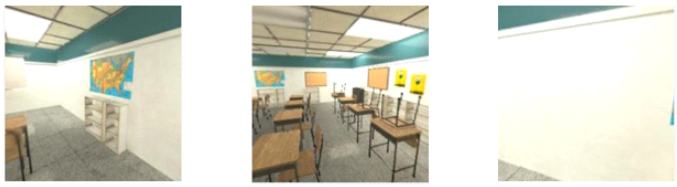
위의 영상을 학습시켜도 영상자체에 정보가 적기 때문에 이들을 입력하여 실내 위치 추정 시, 정확도가 낮을 것으로 예상된다. 오히려 이러한 적은 정보량을 가진 영상들을 분류하기 위해 신경망이 수정되면 정보가 많은 영상들을 분류하는 성능이 떨어질 수 있다. 그러므로 낮은 정보량의 영상을 학습영상에서 제거하는 것을 고려하였다. 이때 영상의 정보량을 측정하기 위한 수단으로 엔트로피(entropy)를 사용하였다.

3.5.2 엔트로피를 이용한 데이터 정제

정보이론에서 엔트로피는 확률 밀도 함수로 변환된 콘텐츠의 불확실성으로 표현되고 이는 콘텐츠의 정보의 양으로 해석된다(안세웅, 2015). 이를 활용하여 영상을 그레이 스케일(gray scale)로 변환하거나 영상의 각 채널에서 엔트로피를 계산하면 그레이 스케일 영상, 혹은 채널별 정보의 양을 정량화 할 수 있다. 영상의 화소값이 이산적일 때, 전체 화소의 수를 n, k값을 가진 화소의 수를 n_k라고 하면 k값을 가진 화소가 등장할 확률 p_k와 영상의 엔트로피 h는 수식(3-3)으로 계산된다.

$$p_k = \frac{n_k}{n}$$
$$h = -\Sigma_i p_i \log_2 p_i \qquad (3\text{-}3)$$

즉 엔트로피는 화소값이 등장할 확률에 로그를 취한값의 기댓값이다. 이 값이 최대가 되는 경우는 영상에 모든 화소값이 동일한 빈도로 등장할 때이며, 최소가 되는 경우는 한 화소값만 영상에 등장할 경우이다. 즉, 소수의 화소값들에 등장 빈도가 편중 될수록 엔트로피는 낮아지게 된다. 다음 [그림 3-17]과 [표 3-2]는 영상과 엔트로피의 관계를 보여주는 예시이다.



[그림 3-17] 구역 1의 원근투영 영상 예시
(a),(b) 엔트로피가 상대적으로 높은 영상
(c) 엔트로피가 상대적으로 낮은 영상

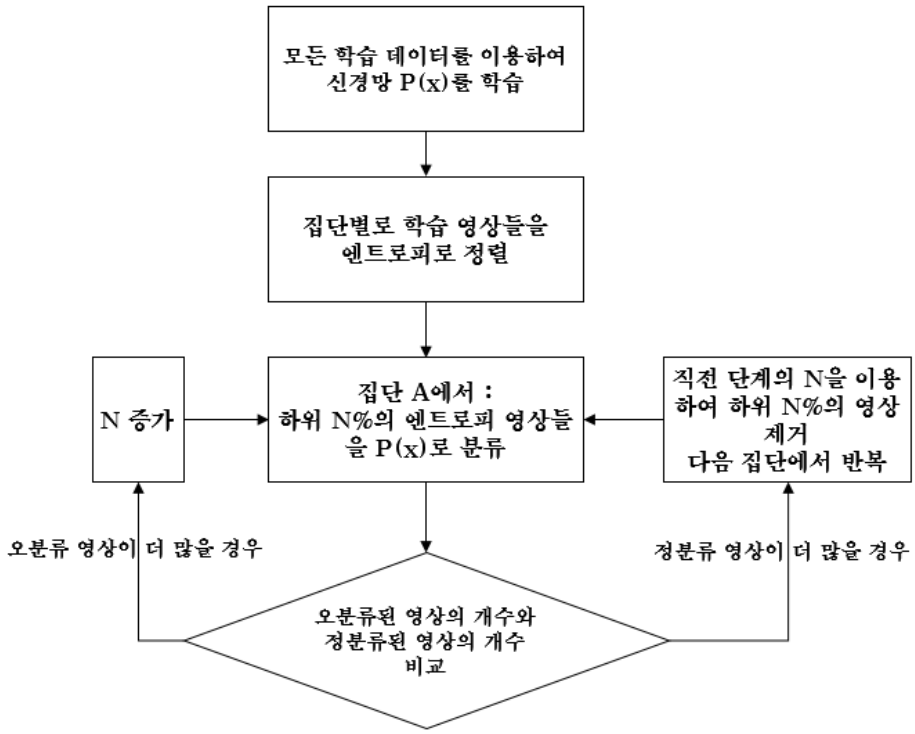
[표 3-2] [그림 3-17]의 채널별 엔트로피

그림 / 채널	(a)	(b)	(c)
R	7.1848	7.5334	4.0722
G	7.1164	7.4017	4.1396
B	7.3108	7.4156	5.7040

[그림 3-17] (c)의 영상은 빈 벽을 표현하기 때문에 (a)와 (b)의 영상에 비해 등장하는 화소값들이 편중되어 있음을 알 수 있다. 그러므로 엔트로피는 전 채널에 걸쳐 두 영상보다 낮은 것을 확인 할 수 있다.

본 연구에서 다중 채널을 가진 영상의 엔트로피는 R, G, B 각각의 채널 엔트로피를 계산한 후, 셋 중 최댓값으로 정의하였다. 특정 채널의 정보량이 높으면 나머지 채널의 정보량이 낮더라도 정보량이 높은 채널로 인해 실내 위치 추정 정확도가 높을 것으로 판단하였기 때문이다.

그러므로 정보가 부족한 데이터를 제거하기 위해 영상들의 엔트로피를 계산하고 엔트로피가 낮은 순서대로 영상을 제거하는 방법을 제안 할 수 있다. 하지만 제거하는 기준이 되는 엔트로피를 정하는 방법이 우선 필요하다. 이를 위하여 [그림 3-18]과 같은 순서도의 방법을 제안한다.



[그림 3-18] 데이터 정제 순서도

엔트로피가 일정 기준보다 낮으면 정보가 부족하여 오분류가 많아지며 엔트로피가 높은 영상일수록 충분한 정보가 제공되어 오분류가 적어진다고 가정한다. 먼저 모든 학습 데이터들을 이용하여 신경망 P(x)를 학습한다. 여기서 x는 영상이고 P(x)의 출력은 x가 각 집단별로 속할 확률이다. 그리고 각 집단별로 학습 영상들의 엔트로피를 구한 후 그 값으로 정렬한다. 다음으로 N값을 정하고 엔트로피 값을 기준으로 하위 N%의 영상들을 P(x)로 분류한다. 이때 P(x)의 가장 높은 확률값을 갖는 집단

이 x가 속한 집단이면 정분류, 그렇지 않을 경우 오분류로 판단한다. 그 후, 하위 N% 엔트로피 영상들의 정분류 수와 오분류 수를 비교한다. 이 때 오분류 영상이 더 많은 경우, 제거할 비율을 더 키워도 된다고 판단하여 N을 증가시키고 정분류의 빈도가 더 많아질 때 까지 반복한다. 단, 너무 많은 학습 영상이 제거되면 학습영상이 부족할 수 있으므로 N의 상한선을 20%로 두었다. 반복하는 도중 정분류의 빈도가 더 많아질 경우, 직전의 N값을 이용하여 하위 N%의 영상들을 제거한다. 제거 후 모든 집단에서 데이터 정제가 완료될 때까지 다음 집단에서 같은 과정을 반복한다. 본 실험에서는 초기 N을 1로 설정하고 N의 증가분을 1로 설정하였다.

3.6 분석 방법

분석은 시험 데이터로 분류 정확도를 비교하고 학습 데이터로 신경망을 학습하는데 소모된 시간을 비교하는 방법으로 진행되었다. 시험 데이터로 사용할 영상은 획득된 원근 투영 영상의 일부를 사용하였다. 분류 정확도를 분석하기 위해 confusion matrix를 구하고 구역별로 정확도가 비교적 낮거나 높은 이유를 분석하였다. 분류 정확도는 top-1, top-3 정확도를 사용하였다. top-1 정확도를 계산할 때 , 가장 높은 속할 확률을 가진 집단과 입력된 영상이 속한 집단과 동일할 경우, 올바른 장소를 인식하였다고 판단하였다. 그리고 top-3 정확도를 계산할 때는 영상이 속할 확률이 첫 번째, 두 번째 그리고 세 번째로 높은 집단 중 입력된 영상이 속한 집단과 동일한 집단이 있을 경우, 올바른 장소를 인식하였다고 판단하였다. 정확도는 (올바른 장소 인식 수)/(전체 영상 수)를 사용하였다.

어떤 분류기가 다른 분류기보다 약간의 성능 우위가 있어도 학습이나 분류에 사용되는 시간이 다른 분류기에 비해 지나치게 높으면 효율적인 분류기라고 할 수 없다. 그러므로 신경망의 학습에 걸리는 시간과 분류 시간 또한 측정하였다.

4. 실험결과 및 분석

4.1 영상 데이터베이스 구축 결과

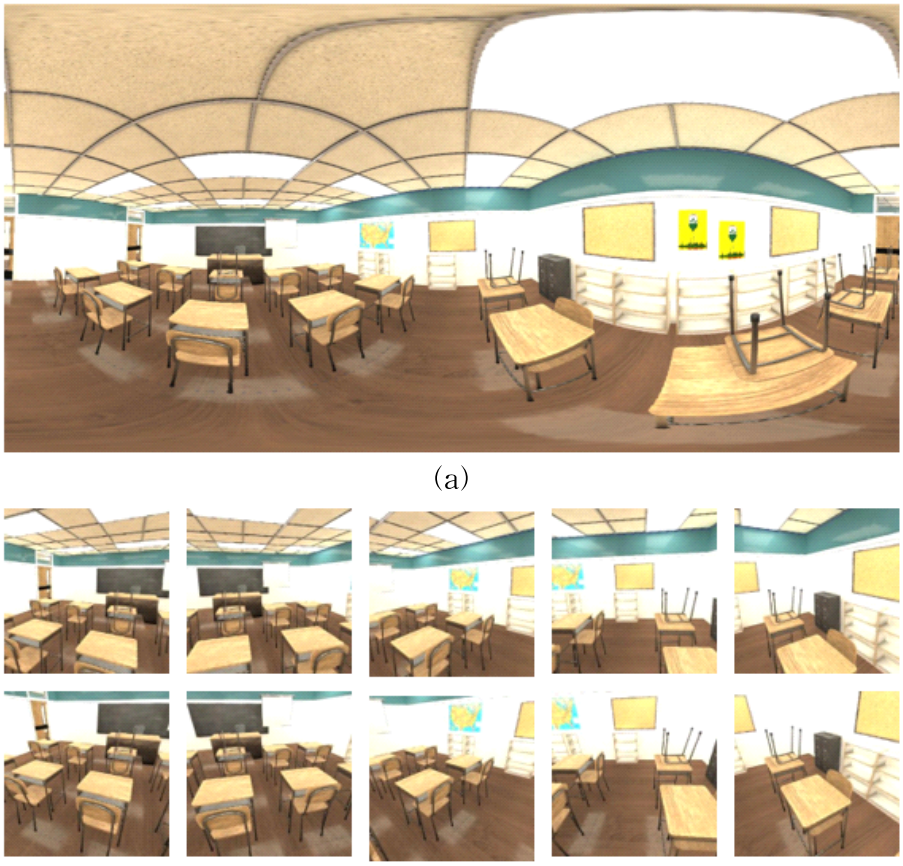
본 연구에서는 대량의 영상을 효율적으로 취득하기 위하여 전방향 영상을 촬영한 후, 이를 원근 투영 영상으로 변환하는 방법을 사용하였다. [표 4-1]은 그 결과를 나타내며 영상 데이터베이스 구축을 위한 설정값들과 생성된 영상들의 개수를 보여준다.

[표 4-1] 영상 데이터베이스 구축에 필요한 설정값 및 구축 결과

항목	값
구역 당 전방향 영상의 개수	100개
전방향 영상의 총 개수	1,500개
원근 투영 영상 시야각	수평, 수직 모두 70°
수평각(φ)의 범위 및 간격	0°~360°, 36° 간격
종 중복도	48.5%
수직각(θ)의 범위 및 간격	60°~120°, 30° 간격
횡 중복도	51.1%
전방향 영상 1개당 생성된 원근 투영 영상 개수	30개
구역 당 원근 투영 영상 개수	3,000개
총 원근 투영 영상 개수	45,000개

전방향 영상을 원근 투영 영상으로 변환할 때, 시야각은 스마트폰 카메라의 시야각을 고려하여 70°로 설정하였다. 또한 원근 투영 영상을 생성하기 위한 광축의 수평각은 36°의 배수로 설정하였고 수직각은 60°, 90° 그리고 120°로 설정하였다. 결과적으로 구역 당 100개의 전방향 영상에서 3,000개의 원근 투영 영상을 생성하였고 모든 구역에서 총 45,000개의 원근 투영 영상을 생성하였다. [그림 4-1]은 그 예시로써, 취득한

전방향 영상 한 장과 그 한 장으로 생성한 원근 투영 영상들의 일부이다. 모든 구역에서 취득한 전방향 영상의 예시는 부록에 수록하였다. [그림 4-1] (b)의 영상은 각각 핀홀 카메라가 같은 위치에서 수평각 36°, 수직각 30°만큼 회전한 후 촬영된 영상으로 볼 수 있다.



(a)

(b)

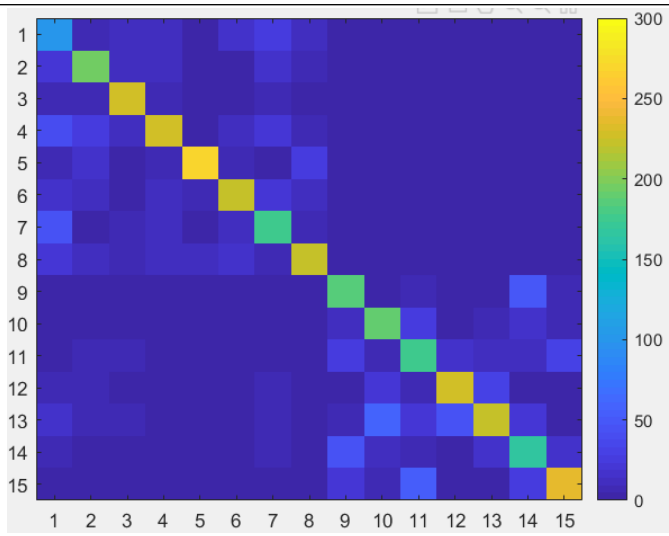
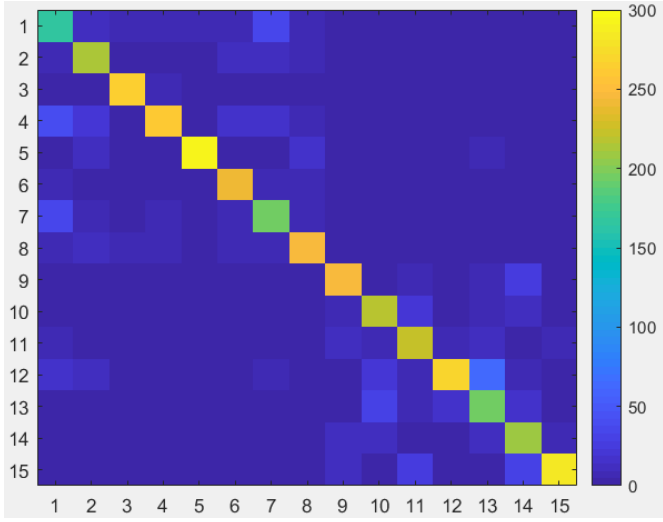
[그림 4-1] 구역 1의 전방향 영상과
전방향 영상으로 생성된 원근 투영 영상의 예시
(a) 구역 1 전방향 영상
(b) 해당 전방향 영상으로부터 생성된 원근 투영 영상의 일부

4.2 분류 결과 분석

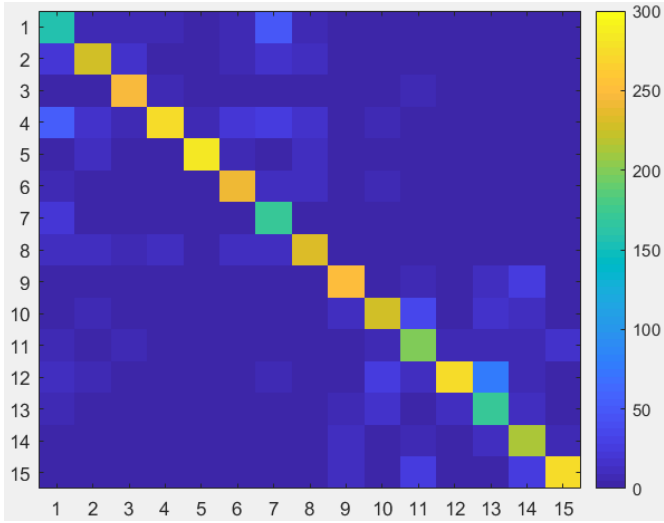
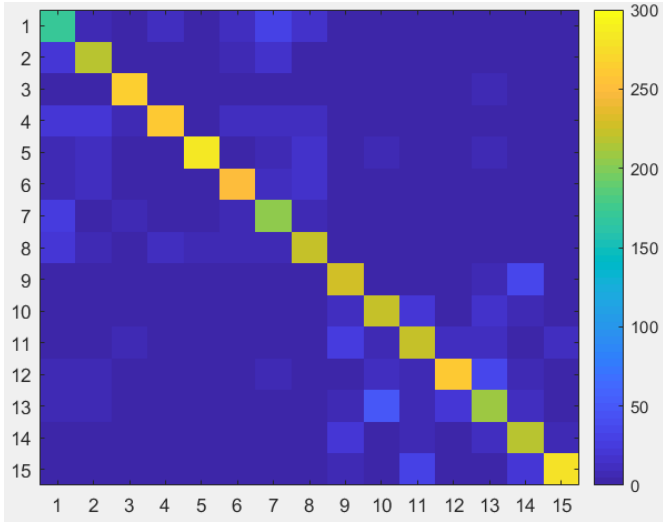
다음 [표 4-2]에서 신경망 종류별 그리고 전이학습 유형별로 실내 위치 추정의 평가를 위한 시험 데이터의 분류 confusion matrix, 시험 데이터 분류 top-1, top-3 정확도, 그리고 학습과 시험 데이터 분류 시 소요된 시간을 정리하였다. 시험 데이터는 각 구역별 원근 투영 영상 300개로 이들은 학습 데이터 및 검증 데이터가 취득된 위치와 다른 곳에서 촬영된 영상으로 구성되어 있다. 이 표에서 confusion matrix의 열은 실제 속한 구역을 나타내고 행은 신경망으로 분류한 구역을 나타낸다. 즉, i번째 행, j번째 열의 요소 a_{ij}는 실제는 j구역이고 신경망으로 분류 시 i구역으로 분류 된 시험 영상의 개수를 뜻한다. Confusion matrix는 15개 집단을 나타내기 위해서 15×15의 행렬로 구성되어야 하지만 15×15행렬을 본문에 삽입할 경우 숫자가 과도하게 많아 데이터 파악이 쉽지 않다. 그러므로 데이터 경향성을 원활하게 파악하기 위해 숫자로 표현된 confusion matrix 대신 숫자에 색상을 부여하여 confusion matrix를 시각화하였다. 이때 시각화된 confusion matrix의 오른쪽에 색상이 의미하는 숫자를 알 수 있도록 색상막대를 추가하였다. 숫자로 표현된 confusion matrix는 부록에 수록하였다. 마지막으로 학습시간은 설정된 조건으로 신경망을 학습할 때, 소요되는 시간이고 분류시간은 시험 데이터 한 장을 입력받고 분류까지 소모되는 평균 시간이다.

[표 4-2] 신경망 종류별, 전이학습 유형별 학습 후 평가 결과

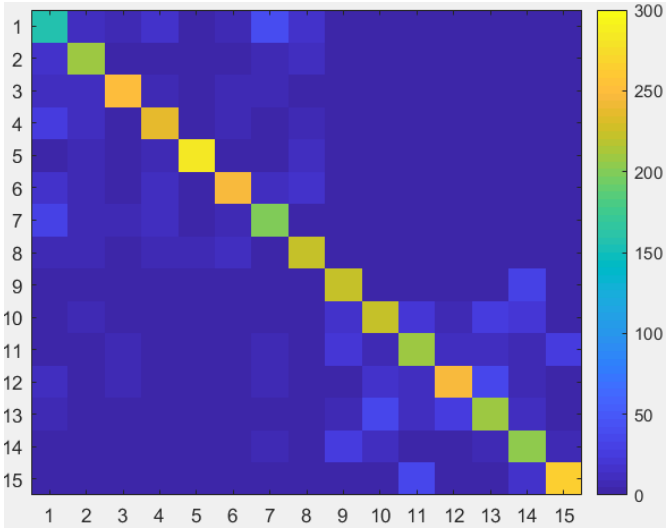
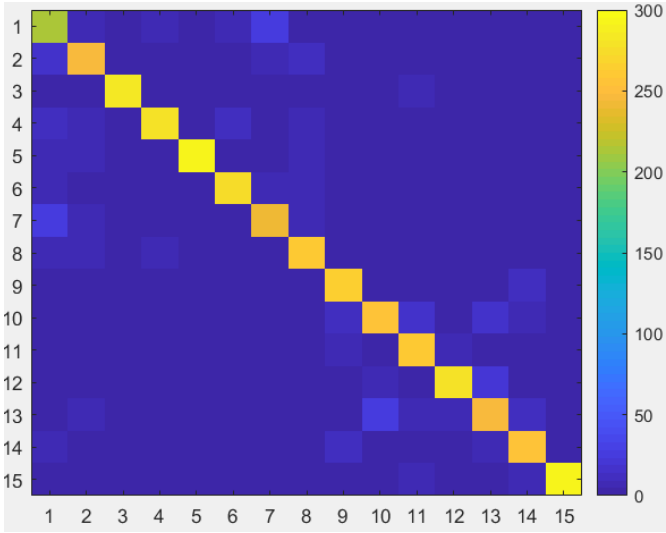
(a) AlexNet의 평가 결과

유형1	 top-1 정확도 : 67.91%, top-3 정확도 : 90.62% 학습시간 : 3분 20초, 분류 시간 : 0.006초
유형3	 top-1 정확도 : 75.58%, top-3 정확도 : 93.69% 학습시간 : 4분 42초, 분류시간 : 0.006초

(b) MobileNet V2의 평가 결과

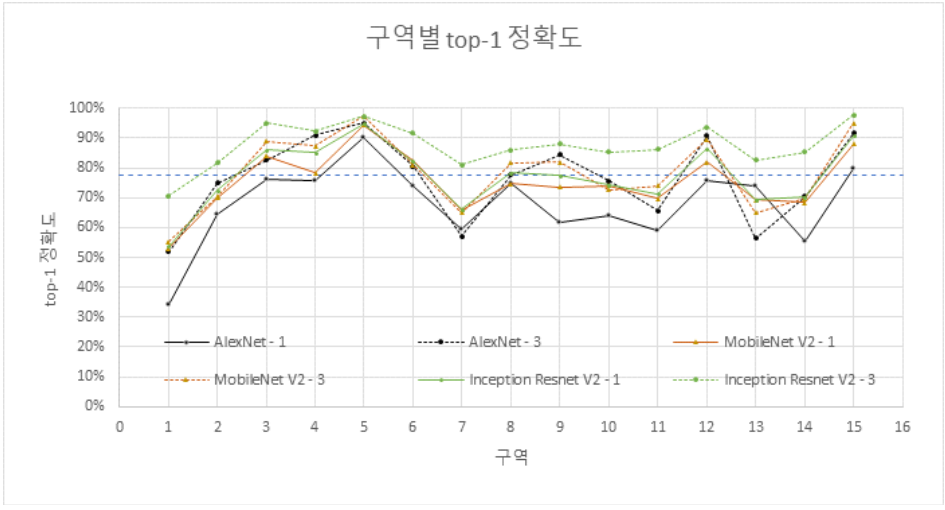
유형1	 top-1 정확도 : 76.56%, top-3 정확도 : 94.13% 학습시간 : 10분 19초, 분류 시간 : 0.033초
유형3	 top-1 정확도 : 78.11% top-3 정확도 : 94.47% 학습시간 : 18분 39초, 분류 시간 : 0.033초

(c) Inception-ResNet V2의 평가 결과

유형1	 top-1 정확도 : 78.56%, top-3 정확도 : 94.76% 학습시간 : 45분 35초, 분류 시간 : 0.127초
유형3	 top-1 정확도 : 87.69%, top-3 정확도 : 97.52% 학습시간 : 72분 21초, 분류 시간 : 0.127초

4.2.1 각 구역별 분류 결과 분석

분류 결과를 정밀하게 분석하기 위해 [그림 4-2]와 같이 구역별로 정확도를 계산하였다. 이 그림에서 세로축은 top-1 정확도이며 파란색 점선은 모든 합성곱 신경망의 평균 top-1 정확도이다. 또한 범례는 (신경망 종류)-(전이학습 유형)으로 표시하였다.

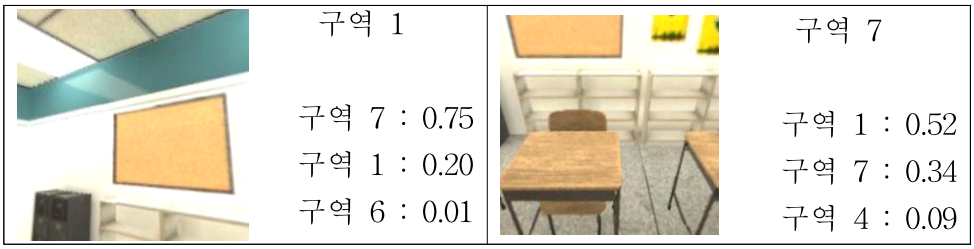


[그림 4-2] 신경망 종류별, 전이학습 유형별 학습 후 구역별 정확도

[그림 4-2]의 구역별 top-1 정확도는 모든 신경망과 전이학습 유형에서 비슷한 경향을 보인다. 특히 정확도가 비교적 낮은 구역은 교실인 구역 1, 구역 7과 복도인 구역 11, 구역 13이고 정확도가 비교적 높은 구역은 교실인 구역 3, 구역 5와 복도인 구역 12, 구역 15임을 알 수 있다.

구역 1과 구역 7의 정확도가 낮은 이유를 알기 위해 confusion matrix를 보면 모든 신경망과 전이학습 유형에서 a_{17}과 a_{71}이 상대적으로 높음을 알 수 있다. 이렇게 오분류 된 이유는 구역 1과 구역 7의 차이점은 책상과 의자의 개수뿐이기 때문이다. 그러므로 구역 1과 구역 7의 영상들 중 책상과 의자의 개수가 소수이거나 혹은 존재하지 않는 영상을 입력한 경우, 두 장소의 차이점인 책상과 의자의 개수를 분별할 수 없으므

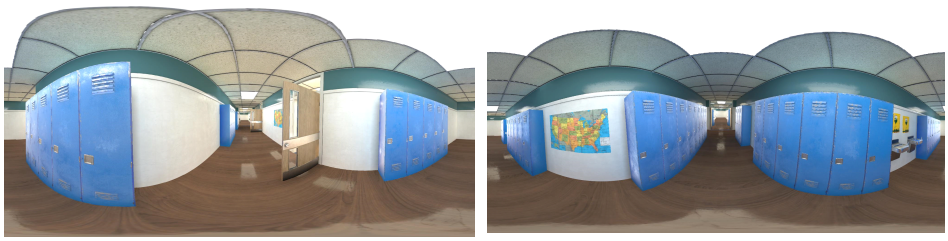
로 정확한 분류가 이루어지지 않았다고 할 수 있다. 그 예시로 다음 [그림 4-3]은 구역 1영상을 구역 7로, 구역 7영상을 구역 1로 오분류한 영상의 예시이며 그 영상이 속할 확률이 가장 높은 3개의 구역과 평균 확률을 표시하였다. 그림과 같이 구역 1과 구역 7의 영상들은 상호간으로 분류한 빈도가 높았기 때문에 분류 정확도가 낮았다.



(a) (b)

[그림 4-3] 구역 1과 구역 7의 오분류 된 시험 영상과 결과 확률예시
(a) 구역7로 오분류 된 구역 1의 영상
(b) 구역1로 오분류 된 구역 7의 영상

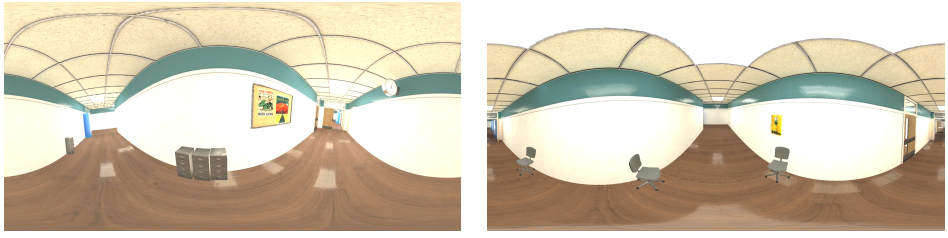
다음으로 confusion matrix에서 복도인 구역 11은 구역 15로 오분류된 빈도가 높음을 알 수 있다. 그 이유를 알기 위해 구역 11과 구역 15를 영상을 살펴보면 [그림 4-4]와 같이 구역 11은 소수의 사물함이 존재하고 구역 15는 사물함으로 가득찬 복도임을 알 수 있다. 이 사물함들은 같은 종류이기 때문에 신경망이 구역 11과 구역 15의 사물함을 구분하지 못한 것으로 해석할 수 있다. 또한 구역 15는 사물함이 모든 구역 중 가장 많이 존재한다. 그러므로 사물함이 포함된 영상들은 구역 15일 가능성이 가장 높고 이러한 특징을 가진 영상들로 신경망을 학습시켰기 때문에 복도와 사물함이 함께 촬영된 영상은 구역 15의 영상으로 분류될 확률이 증가하였다. 결과적으로 구역 11의 사물함 영상은 구역 15로 분류되었을 가능성이 크기 때문에 분류 정확도가 낮았다고 할 수 있다.



(a) (b)

[그림 4-4] 같은 종류의 사물함이 설치된
구역 11과 구역 15의 전방향 영상
(a) 구역 11 (b) 구역 15

다음은 구역 13에 대한 설명이다. confusion matrix를 보면 구역 13은 구역 12로 오분류 된 빈도가 높다. 그 이유는 [그림 4-4]에서 볼 수 있듯, 구역 13과 구역 12가 모두 '비어있는 복도'라는 공통점을 가지는 장소이기 때문이다. 앞에서 언급한 구역 11이 정확도가 낮은 이유와 마찬가지로 구역 12에는 구역 13보다 존재하는 물체가 더 적기 때문에 신경망이 아무것도 없는 복도의 영상을 구역 12로 분류할 확률이 높게 학습되었음을 판단할 수 있다. 그러므로 구역 13의 물체가 없는 복도 영상들은 높은 확률로 구역 12로 오분류가 되어 분류 정확도가 낮음을 추론할 수 있다.

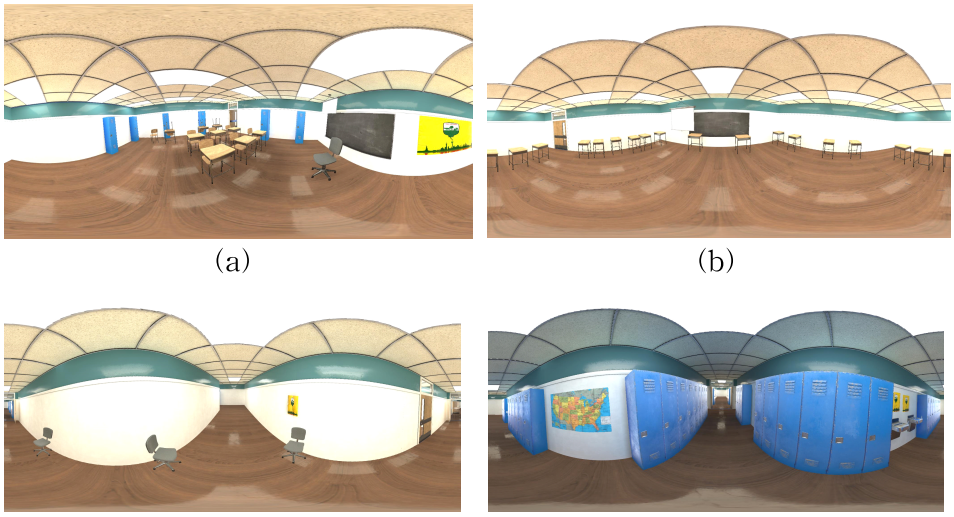


(a) (b)

[그림 4-5] 비어있는 복도라는 공통인 구역 13과 구역 12의 전방향 영상
(a) 구역 13 (b) 구역 12

정확도가 비교적 높은 구역인 구역 3, 구역 7, 구역 12 그리고 구역 15는 모두 구역 고유의 특징이 강한 장소라고 할 수 있다. 즉, [그림 4-6]과 같이 구역 3은 교실들 중에서 유일하게 사물함이 존재하는 교실, 구

역 5는 가구가 적은 비어있는 교실, 그리고 구역 12와 구역 15는 각각 물체가 없다는 특성과 사물함이 존재한다는 특성이 가장 잘 나타나는 장소이다. 이렇게 영상에 잘 나타나는 특성을 가지는 장소에서 학습 영상을 취득 시, 그 특성이 드러나는 영상이 취득되는 빈도가 높다. 앞에서 언급하였듯, 신경망은 비슷한 특징을 가진 영상들이 학습 영사에 있으면 그 특징이 출현하는 빈도가 가장 높은 장소로 분류되도록 학습된다. 그러므로 신경망으로 시각적 실내 위치 추정 시, 위의 결과처럼 특징이 강한 장소들은 분류 정확도가 비교적 높은 결과를 보인다.



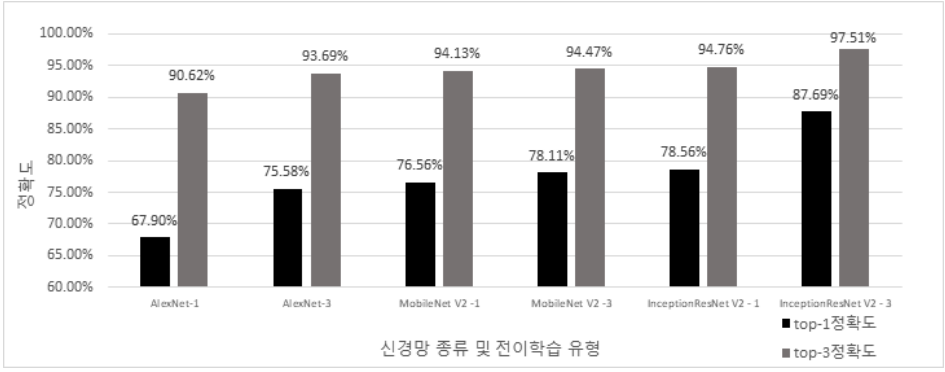
(a) (b)

(c) (d)

[그림 4-6] 강한 특징으로 인해 정확도가 높은 구역
(a) 구역 3 (b) 구역 5 (c) 구역 12 (d) 구역 15

4.2.2 합성곱 신경망 종류별, 전이학습 유형별 평가

[그림 4-7]은 신경망 종류 및 전이학습 유형에 따라 학습된 신경망으로 시각적 실내 위치 추정의 top-1 정확도 그래프와 top-3 정확도 그래프를 보여준다.



[그림 4-7] 신경망 종류 및 전이학습 유형별
top-1, top-3 정확도 그래프

합성곱 신경망 종류별 실내 위치 추정 성능의 평가를 위하여 AlexNet, MobileNet V2 그리고 Inception-ResNet V2의 top-1, top-3 정확도를 살펴보면 전이학습 유형 1과 유형 3에서 모두 Inception-ResNet V2가 가장 높았으며 AlexNet이 가장 낮았다. 이 정확도 순위는 ImageNet의 영상들을 이용한 분류 성능의 순위와 같다. 이는 복잡성이 높은 신경망이 사물 분류에서 성능이 좋은 이유와 같은 이유로, 신경망 모델이 더욱 복잡하고 깊어짐에 따라 추출할 수 있는 특징이 다양해지기 때문으로 해석된다. 그러므로 ImageNet의 영상들을 이용하여 학습 후 분류 정확도가 높은 모델은 ImageNet의 영상 분류와 비슷한 업무인 시각적 실내 위치 추정 시에도 높은 정확도를 보임을 알 수 있다.

다음으로 전이학습 유형별 비교를 위하여 각 합성곱 신경망 종류별 유형 1과 유형 3의 top-1 정확도와 top-3 정확도를 비교하였다. AlexNet,

MobileNet V2 그리고 Inception ResNet V2에서 전역 계층만을 학습시키는 전이학습 유형 1보다 신경망의 입력층과 가까운 합성곱 계층을 제외한 전 구간을 미세조정하는 전이학습 유형 3으로 학습한 신경망들의 정확도가 평균 6.12% 더 높음을 알 수 있다. 이는 유형 3으로 학습 시, 신경망의 합성곱 계층 필터들이 본 실험의 영상들을 분류할 수 있는 특징을 추출할 수 있도록 수정되기 때문으로 해석된다. 그러므로 더 좋은 실내 위치 추정 성능을 위해 합성곱 계층까지 학습시키는 유형 3의 방법이 더 적합함을 실험적으로 확인할 수 있다.

종합하면 ImageNet의 영상 분류에서 좋은 성능을 발휘하는 합성곱 신경망일수록 시각적 실내 위치 추정에도 좋은 성능을 발휘하며, 그 성능은 전역 연결 계층만을 학습시키는 전이학습 유형 1보다 전역 연결 계층과 합성곱 계층의 일부분을 학습시키는 전이학습 유형 3을 사용할 때, 향상됨을 알 수 있다. 그러나 좋은 성능을 내는 신경망일수록 계층이 많고 모수가 많으므로 학습시간 및 분류시간이 더욱 길어진다. 때문에 신경망 간 학습시간과 분류시간 차이는 정확도의 차이보다 상대적으로 더욱 크다. AlexNet과 MobileNet V2의 유형 3으로 학습시킬 때, 정확도는 2.5%만큼 차이가 있으나, 학습시간과 분류시간은 4배정도 더 소모된다. 마찬가지로 AlexNet과 Inception-ResNet V2를 유형 3으로 학습 시 정확도는 12.2%의 차이가 있으나 학습시간과 분류시간은 20배의 차이가 발생하므로 요구되는 학습시간과 분류시간을 고려하여 합성곱 신경망을 선택하여야 한다. 또한 top-3 정확도의 경우 모든 신경망이 90%가 넘는 수치를 보인다. 즉, 비슷한 특성을 가진 장소가 본 연구처럼 상대적으로 소수일 경우, 영상이 속한 장소가 신경망의 분류 결과로 도출된 확률값이 가장 높은 3개의 장소 중 한 개일 가능성은 90%이상임을 알 수 있다.

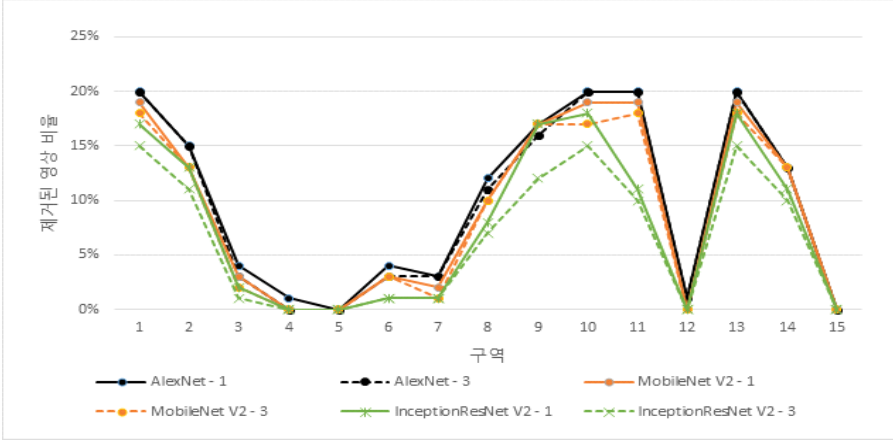
4.3 데이터 정제 평가 결과

4.3에서는 제안된 데이터 정제 방법으로 제거된 데이터 개수를 살펴보고 그 효과를 보여준다. 먼저 다음 [표4-3]에서 3.5의 데이터 정제 방법을 적용하여 합성곱 신경망 별, 전이학습 유형별로 제거된 학습 영상의 개수와 비율을 나타내었다. 또한 검증 영상과 시험 영상도 엔트로피 값으로 정렬 후 같은 비율로 영상들을 정제하였다.

[표 4-3] 신경망 종류, 전이학습 유형별 제거된 학습 영상의 개수와 비율

	유형 1	유형 3
AlexNet	3150(10.00%)	3045(9.67%)
MobileNet V2	2877(9.13%)	2730(8.67%)
Inception-ResNet V2	2457(7.80%)	2058(6.53%)

이때 제거된 학습 영상의 개수와 비율을 살펴보면, 정확도가 높은 경우일수록 제거되는 영상은 적은 경향을 보인다. 더 자세한 분석을 위해 구역별로 제거된 영상의 비율을 계산하였다. 다음 [그림 4-8]은 신경망 종류별과 전이학습 유형별 실내 위치 추정 후, 데이터 정제 방법을 사용하여 구역별로 제거된 영상의 비율을 표현한 그래프이다.



[그림 4-8] 구역별 제거된 영상의 비율 그래프

위와 같이 구역별로 제거된 영상의 비율이 다른 이유는 신경망의 분류 정확도에 따라 제거에 사용될 엔트로피의 기준이 결정되기 때문이다. 데이터 정제 방법에서 정보가 부족한 영상을 제거하기 위해, 정분류 된 영상의 비율이 50%이상이 될 때 까지 제거할 기준이 되는 엔트로피를 증가시켰다. 그러므로 정확도가 높은 구역과 합성곱 신경망 종류, 그리고 전이학습 유형 일수록 정분류의 빈도가 높아 엔트로피의 기준이 낮기 때문에 제거되는 영상이 적음을 알 수 있다. 특히 구역 15의 경우, 엔트로피 값이 낮은 영상들은 사물함이 포함된 영상이 대다수였고 이들의 절반 이상은 구역 15로 정분류되어 제거된 영상의 비율이 모든 경우에서 낮았다. 마찬가지로, 구역 5와 구역 12는 물체가 거의 없는, 비어있는 특징의 장소로 해당 구역의 영상들은 낮은 엔트로피 값을 가지는 빈 벽만 포함한 영상이 다수였으나, 학습된 합성곱 신경망이 그러한 영상들을 구역 5와 구역 12로 분류하도록 학습되었다. 그러므로 엔트로피가 낮은 구역 5와 구역 12의 영상들을 분류 시, 정분류 된 빈도가 높아 제거된 비율이 상대적으로 낮았다.

다음으로 데이터 정제의 효과를 알아보기 위한 데이터 정제 전과 후의 정확도를 살펴본다. [표 4-4]와 [표 4-5]는 각각 정제된 영상들을 이용하여 학습한 합성곱 신경망들과 정제하지 않은 영상들을 이용하여 학습한 합성곱 신경망들의 top-1 정확도, top-3 정확도 그리고 정제 전과 후의 정확도의 차이를 나타낸 표이다.

[표 4-4] 학습 데이터 정제 전 영상을 학습한 신경망의 top-1 정확도, 정제 후 영상을 학습한 신경망의 top-1 정확도와 그 차이

합성곱 신경망 종류	전이학습 유형	정제 전	정제 후	차이
AlexNet	유형 1	67.90%	68.89%	0.99%
	유형 3	75.58%	76.78%	1.20%
MobileNet V2	유형 1	76.56%	79.49%	2.93%
	유형 3	78.11%	80.67%	2.56%
Inception ResNet V2	유형 1	78.56%	81.22%	2.67%
	유형 3	87.69%	89.27%	1.58%
평균값		77.40%	79.39%	1.99%

[표 4-5] 학습 데이터 정제 전 영상을 학습한 신경망의 top-3 정확도, 정제 후 영상을 학습한 신경망의 top-3 정확도와 그 차이

합성곱 신경망 종류	전이학습 유형	정제 전	정제 후	차이
AlexNet	유형 1	90.62%	92.29%	1.67%
	유형 3	93.69%	94.53%	0.84%
MobileNet V2	유형 1	94.13%	95.47%	1.33%
	유형 3	94.47%	96.20%	1.73%
Inception ResNet V2	유형 1	94.76%	96.24%	1.49%
	유형 3	97.51%	98.51%	1.00%
평균값		94.20%	95.54%	1.34%

top-1 정확도는 정제 후 평균 1.99%만큼 증가하였으며 top-3 정확도는 정제 후 평균 1.34%만큼 증가하였다. 정확도 증가의 요인을 분석하기 위해 시험영상 중 낮은 엔트로피를 가지는 영상에 주목하였다. 다음 [그림 4-9]와 [표 4-6]은 시험영상들 중 제거된 영상의 엔트로피보다 낮은 엔트로피를 갖는 영상의 예시와 그 영상들을 합성곱 신경망으로 분류한 정확도이다.

[그림 4-9] 구역별 데이터 정제에 사용된 엔트로피 기준보다 작은 엔트로피를 가진 시험 영상의 예시

[표 4-6] 데이터 정제에 사용된 엔트로피 기준보다 낮은 엔트로피를 가진 시험 영상들의 분류 top-1 정확도

합성곱 신경망 종류	전이학습 유형	정제 전	정제 후
AlexNet	유형 1	29.11%	14.29%
	유형 3	28.11%	11.43%
MobileNet V2	유형 1	27.50%	13.21%
	유형 3	17.68%	9.82%
Inception ResNet V2	유형 1	35.36%	19.29%
	유형 3	22.86%	17.86%
평균값		26.77%	14.32%

[그림 4-9]의 예시 영상과 같이 시험 영상 중에는 정보량이 적어 시각적 실내 위치 추정 방법으로 위치를 확인하기 어려운 영상이 존재한다. 이 영상들을 정제 전 영상으로 학습시킨 신경망과 정제 후 영상으로 학습시킨 신경망을 이용하여 분류 정확도를 측정한 결과는 [표 4-6]과 같다. 이렇게 정보가 부족한 영상을 분류하면 두 신경망 모두 낮으면서 비슷한 정확도가 측정되어야 하나, 정제 전의 영상으로 학습시킨 신경망은 모든 경우에서 정제 후의 영상으로 학습시킨 신경망보다 정확도가 높다. 그러므로 정제 전의 영상으로 학습시킨 신경망은 정보가 부족한 영상도 학습이 된 상태로 볼 수 있으나, 이들의 정확도가 신경망들의 평균 정확도보다 낮으므로 학습의 효과가 적다고 할 수 있다. 또한 정보가 적은 영상을 학습시키지 않은, 정제 후 영상을 이용한 신경망들은 정보가 적은 영상에 대한 정확도가 더 낮다. 대신 정보가 [표 4-7]에서 볼 수 있듯, 적지 않은 영상들에 대해 더욱 분류가 잘 되도록 신경망이 학습되어 모든 경우에서 top-1 정확도는 평균적으로 1.99%만큼 증가하였음을 확인가능하다.

[표 4-7] 데이터 정제에 사용된 엔트로피 기준보다 높은 엔트로피를 가진 시험 영상들의 분류 top-1 정확도

합성곱 신경망 종류	전이학습 유형	정제 전	정제 후
AlexNet	유형 1	72.22%	74.87%
	유형 3	80.66%	82.45%
MobileNet V2	유형 1	81.49%	82.92%
	유형 3	83.95%	84.59%
Inception ResNet V2	유형 1	82.21%	83.57%
	유형 3	92.22%	92.57%
평균값		82.11%	83.33%

5. 결론

시각적 실내 위치 추정은 영상을 이용하여 장소를 분류하는 기법이므로, 영상 분류에 뛰어난 성능을 보이는 심층 합성곱 신경망을 이용하는 방법이 활발하게 연구되며 주목받고 있다. 하지만 심층 합성곱 신경망이 높은 성능을 발휘하기 위해서는 많은 모수와 계층이 필요하고 학습 데이터 개수와 계산 자원이 대량으로 요구되며, 다양한 초모수를 고려해야한다는 이유로 전이학습을 이용하는 경우가 많다. 따라서 본 연구에서는 영상기반 합성곱 신경망을 시각적 실내 위치 추정에 사용하기 위한 전이학습과 관련하여 학습 데이터로 이용할 대량의 영상을 얻는 방법, 신경망 종류와 전이학습 유형별 실내 위치 추정 정확도 분석, 그리고 학습에 악영향을 끼치는 데이터를 제거하는 방법에 초점을 맞추어 연구를 진행하였다.

본 연구의 실험은 가상환경에서 실행되었고 카메라 경로를 결정하였다. 가상환경을 구축할 때, 장소들 간 공통점과 특징을 설정하였고, 카메라 경로를 지정할 때는 인간이 이동할 때의 경로로 가정하여 적절한 높이와 임의의 회전을 고려하였다.

대량의 영상 데이터를 얻는 방법은 전방향 카메라를 이용하는 방법으로 다음과 같이 진행되었다. 먼저 구축된 환경에서 전방향 영상들을 취득하였고, 취득한 전방향 영상 1장당 30장의 원근 투영 영상으로 변환하여 영상 데이터베이스를 구축하였다. 여기서 변환은 전방향 영상의 단위구 모형에서 핀홀 카메라 모델의 요소를 설정한 후 설정된 핀홀 카메라에서 촬영된 원근 투영 영상을 얻는 과정으로 진행되었다.

다음으로 구축된 영상 데이터베이스를 이용하여 ImageNet의 영상들로 사전 학습된 AlexNet, MobileNet V2, 그리고 Inception-ResNet V2를 지식 이전하고 미세조정 하였다. 미세조정 시, 전역 연결 계층만 학습하는 전이학습 유형 1과 전역 연결 계층과 합성곱 계층의 일부도 학습하는 전이학습 유형 3으로 각각 세 모델들을 학습시켜 총 여섯 경우의 학습을 진행하였다. 이때, 분석을 위해 구역별 위치 추정 정확도를 평가한 결과,

공통된 특징을 공유하는 구역들은 낮은 정확도를 보였으며, 명확한 특징을 가지는 구역들은 상대적으로 높은 정확도를 보였다. 신경망별 실내 위치 추정 정확도는 Inception-ResNet V2, MobileNet V2 그리고 AlexNet 순으로 높게 측정되었다. 즉, ImageNet의 영상을 잘 분류하는 신경망 일수록 실내 위치 추정에서 좋은 성능을 보였다. 또한 전역 연결 계층만을 학습시키는 전이학습 유형 1보다 일부 합성곱 계층까지 학습시키는 전이학습 유형 3으로 학습시킬 때 정확도가 평균적으로 6.12%만큼 향상됨을 확인하였다. 그러므로 전이학습 유형 3으로 학습시킬 때, 실내 위치 추정 성능이 향상됨을 확인하였다. 하지만 더 복잡한 구조를 가질수록, 정확도가 높아지는 것에 비해 학습시간과 분류시간이 더욱 증가함을 알 수 있었다. 즉, MobileNet V2와 Inception-ResNet V2는 AlexNet과 비교하여 각각 2.5%와 12.2%의 정확도 향상을 보였으나, 학습시간과 분류시간은 각각 4배와 20배만큼 더 소모되었다.

다음으로 생성된 원근 투영 영상을 살펴보던 중, 정보량이 적은 영상들을 관찰하였고 이들을 제거하는 데이터 정제를 고려하였다. 정보량이 적은 영상들이 포함된 데이터베이스를 신경망에 학습 시킨 후 분류할 경우, 상대적으로 낮은 정확도를 보이기 때문에 이들을 제거하는 것이 오히려 신경망 학습에 도움이 될 것으로 판단하였다. 데이터 정제를 위해 영상들의 정보량을 엔트로피로 수치화 하였고, 엔트로피가 낮은 영상들은 정보가 부족해 신경망으로 분류가 어렵다는 특징을 이용하여 제거할 영상의 엔트로피를 결정하였다. 결과적으로, 정보가 부족한 영상을 제거한 데이터베이스로 학습된 신경망들은 정보량이 낮지 않은 영상들을 더 잘 분류하기 때문에 정제 전 영상 데이터베이스로 학습된 신경망들보다 실내 위치 추정 정확도가 평균적으로 1.99%만큼 향상되었음을 확인 할 수 있었다.

이상의 연구는 선행 연구들처럼 초기 신경망을 사용하는 것보다는 발전된 최신의 신경망을 사용하고 전이학습 유형 3을 사용하는 것이 실내 위치 추정에서 유리함을 실험적으로 증명했다는 점에서 가치가 있다. 또한 대량의 영상을 얻기 위하여 전방향 영상을 사용하는 방법을 제안하였

고 이는 대량의 영상이 필요한 다른 분야에 사용될 수 있다. 마지막으로, 정보가 부족한 영상을 제거하는 데이터 정제는 분류 정확도를 향상시켰고 이 방법은 다른 기계학습의 사전작업에도 사용될 수 있음에 그 가치가 있다.

본 연구는 실제세계가 아닌 가상환경에서 이루어졌다는 한계가 있다. 그러므로 본 논문에서 사용한 전방향 영상과 합성곱 신경망을 이용한 실내 위치 추정은 실제세계에서 발생할 수 있는 물체의 이동으로 인한 영향을 고려해야하고, 전방향 카메라의 잡음과 왜곡을 보정하는 과정이 선행된 다음에 사용가능하다. 또한 실험에서 사용한 위치 추정기법은 합성곱 신경망 기반이므로 신경망이 가지는 한계를 그대로 가진다. 신경망은 학습시킨 데이터의 범위 안에 존재하는 데이터만 예측이 가능하다는 문제점이 있어 학습 시키지 않은 다른 장소의 영상을 입력하였을 때, 다른 장소임을 판단할 수 없다는 한계를 지닌다. 따라서 추후 실제세계에서의 적용과 학습시키지 않은 장소를 판단하는 방법을 연구하여 더욱 범용적이고 안정적인 실내 위치 추정 시스템을 구축하고자 한다.

참 고 문 헌

Aghayari, S., Saadatseresht, M., Omidalizarandi, M., & Neumann, I. (2017). Geometric Calibration of Full Spherical Panoramic Ricoh-Theta Camera. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 237-245.

Alahi, A., Ortiz, R., & Vandergheynst, P. (2012, June). Freak: Fast retina keypoint. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, 510-517.

Ali, M., Sahin, F., Kumar, S., & Savur, C. (2017, October). 360° view camera based visual assistive technology for contextual scene information. In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2135-2140.

Badino, H., Huber, D., & Kanade, T. (2012, May). Real-time topometric localization. In 2012 IEEE International Conference on Robotics and Automation, 1635-1642.

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (SURF). Computer vision and image understanding, 110(3), 346-359.

Bay, H., Tuytelaars, T., & Van Gool, L. (2006, May). Surf: Speeded up robust features. In European conference on computer vision, 404-417.

Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., & Fua, P. (2011). BRIEF: Computing a local binary descriptor very fast. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(7), 1281-1298.

Calonder, M., Lepetit, V., Strecha, C., & Fua, P. (2010, September). Brief: Binary robust independent elementary features. In European conference on computer vision, 778-792.

Caruana, R., Lawrence, S., & Giles, C. L. (2001). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In Advances in neural information processing systems, pp. 402-408.

Chen, Y., Chen, R., Liu, M., Xiao, A., Wu, D., & Zhao, S. (2018). Indoor Visual Positioning Aided by CNN-Based Image Retrieval: Training-Free, 3D Modeling-Free. Sensors, 18(8), 2692.

Fei-Fei, L., & Perona, P. (2005, June). A bayesian hierarchical model for learning natural scene categories. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05),Vol 2, 524-531.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A.,& He, K. (2017). Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677.

Ha, I., Kim, H., Park, S., & Kim, H. (2018). Image retrieval using BIM and features from pretrained VGG network for indoor localization. Building and Environment, 140, 23-31.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778.

Ho, K. L., & Newman, P. (2007). Detecting loop closure with scene sequences. International Journal of Computer Vision, 74(3), 261-286.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 4700-4708.

Kim, D. K., & Chen, T. (2015). Deep neural network for real-time autonomous indoor navigation. arXiv preprint arXiv:1511.04668.

Kornblith, S., Shlens, J., & Le, Q. V. (2018). Do better imagenet models transfer better?. arXiv preprint arXiv:1805.08974.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, 1097-1105.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436.

Leutenegger, S., Chli, M., & Siegwart, R. (2011). BRISK: Binary robust invariant scalable keypoints. In 2011 IEEE international conference on computer vision (ICCV), 2548-2555.

Lowe, D. G. (1999, September). Object recognition from local scale-invariant features. In iccv, Vol. 99, No. 2, 1150-1157.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.

Lowry, S., Sünderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P., & Milford, M. J. (2015). Visual place recognition: A survey. IEEE Transactions on Robotics, 32(1), 1-19.

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013, June). Rectifier nonlinearities improve neural network acoustic models. In Proc. icml, Vol. 30, No. 1, 3.

Mathworks, Pretrained Deep Neural Networks, 검색일 6월 5일, 2019년, https://kr.mathworks.com/help/deeplearning/ug/pretrained-convolutional-neural-networks.html

Mikolajczyk, K., & Schmid, C. (2001, July). Indexing based on scale invariant interest points. In International Conference on Computer Vision (ICCV'01), Vol. 1, 525-531.

Oh, J., & Lee, B. (2019). Condition-invariant Place Recognition Using Deep Convolutional Auto-encoder. Journal of Korea Robotics Society, 14(1), 8-13.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision, 42(3), 145-175.

Rosten, E., & Drummond, T. (2006, May). Machine learning for high-speed corner detection. In European conference on computer vision, 430-443.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. R. (2011, November). ORB: An efficient alternative to SIFT or SURF. In ICCV, Vol. 11, No. 1, 2.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4510-4520.

Scaramuzza, D. (2014). Omnidirectional camera. Computer Vision: A Reference Guide, 552-560.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sivic, J., & Zisserman, A. (2003, October). Video Google: A text retrieval approach to object matching in videos. 1470.

Specht, D. F. (1990). Probabilistic neural networks. Neural networks, 3(1), 109-118.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), 1929-1958.

Sünderhauf, N., Dayoub, F., Shirazi, S., Upcroft, B., & Milford, M. (2015). On the performance of convnet features for place recognition. arXiv preprint arXiv:1501.04158.

Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., & Milford, M. (2015). Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. Proceedings of Robotics: Science and Systems XII.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-First AAAI Conference on Artificial Intelligence.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1-9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2818-2826.

Werner, M., Hahn, C., & Schauer, L. (2016, October). DeepMoVIPS: Visual indoor positioning using transfer learning. In 2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 1-7.

Werner, M., Kessel, M., & Marouane, C. (2011, September). Indoor positioning using smartphone camera. In 2011 International Conference on Indoor Positioning and Indoor Navigation, 1-6.

Zhang, F., Duarte, F., Ma, R., Milioris, D., Lin, H., & Ratti, C. (2016). Indoor space recognition using deep convolutional neural network: a case study at MIT campus. arXiv preprint arXiv:1610.02414.

Zhu, C. (2017). Place recognition: An overview of vision perspective. arXiv preprint arXiv:1707.03470.

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 8697-8710

강대기. (2016). 딥 러닝 기반 기계학습 기술 동향. 정보통신기술진흥센터 주간기술동향, 12-24.

안세웅, & 이상훈. (2015). 엔트로피 관점에서 2D 와 3D 동영상의 시각적 정보량 측정방법. 한국방송미디어공학회 학술발표대회 논문집, 8-10.

양윤모. (2004). 전방향 비전센서 및 응용. 한국정보과학회 학술발표논문집, 117-122.

오정현. (2018). Sequence-based place recognition using deep learning features in changing environments (Doctoral dissertation, 서울대학교 대학원).

유재준. (2013). 실내 위치기반 서비스 기술 및 서비스 개발 동향. 주간기술동향 (정보통신산업진흥원), 14-26.

이한수, 김종근, 유정원, 정영상, & 김성신. (2018). 전이학습 기반의 합성곱 신경망을 이용한 다중클래스 분류에 관한 연구. 한국지능시스템학회 논문지, 28(6), 531-537.

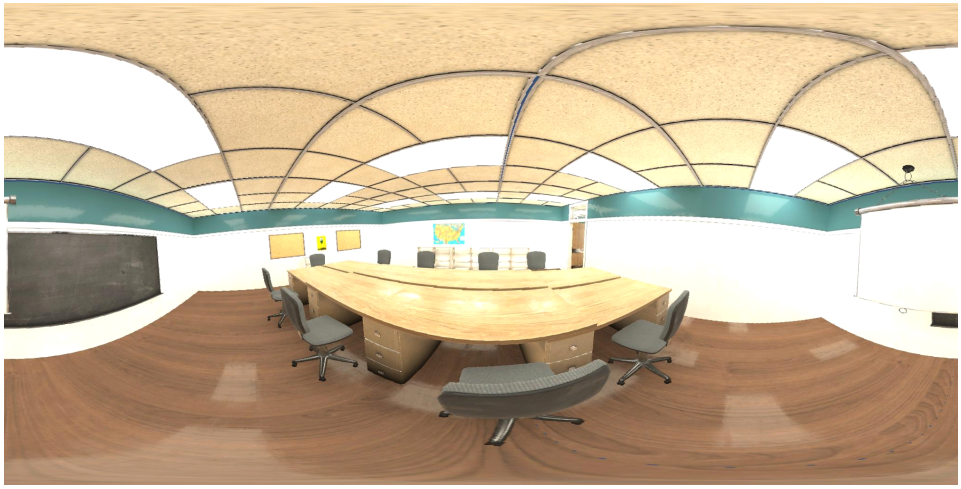
추민곤. (2019). 자료기반 다중모달 합성곱 신경망을 사용한 추가정의 최적 위치 선정 연구 (Doctoral dissertation, 서울대학교 대학원).

한국인터넷진흥원.(2017) 국내외 LBS 산업 동향 보고서, KISA REPORT.

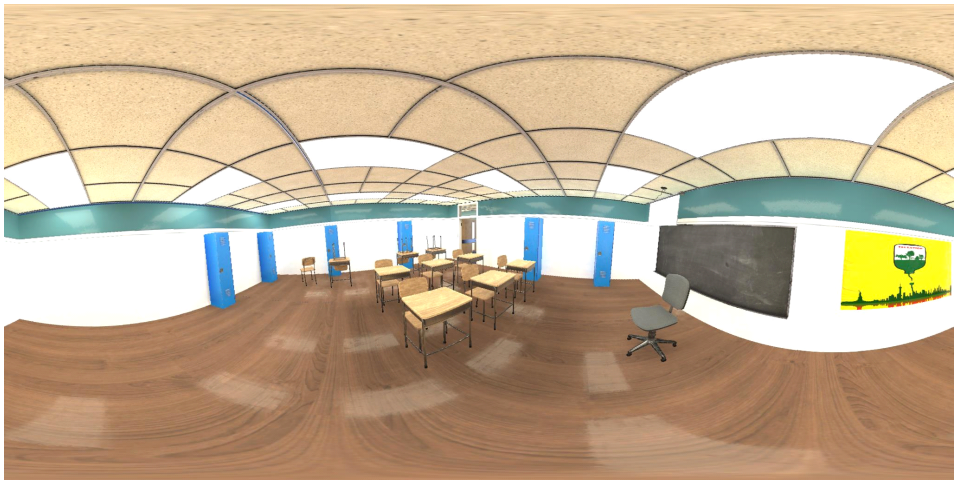
부록

A.1 구역별 전방향 영상

[그림 A.1-1] 구역 1의 전방향 영상

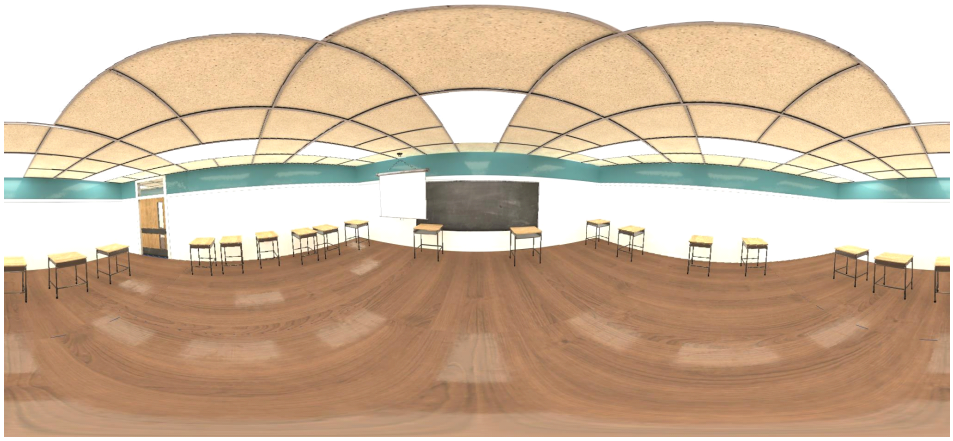


[그림 A.1-2] 구역2의 전방향 영상

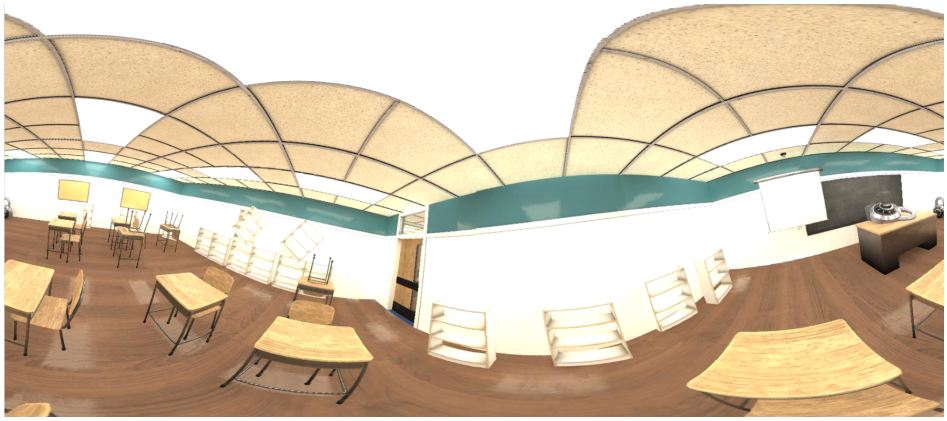


[그림 A.1-3] 구역3의 전방향 영상

[그림 A.1-4] 구역4의 전방향 영상

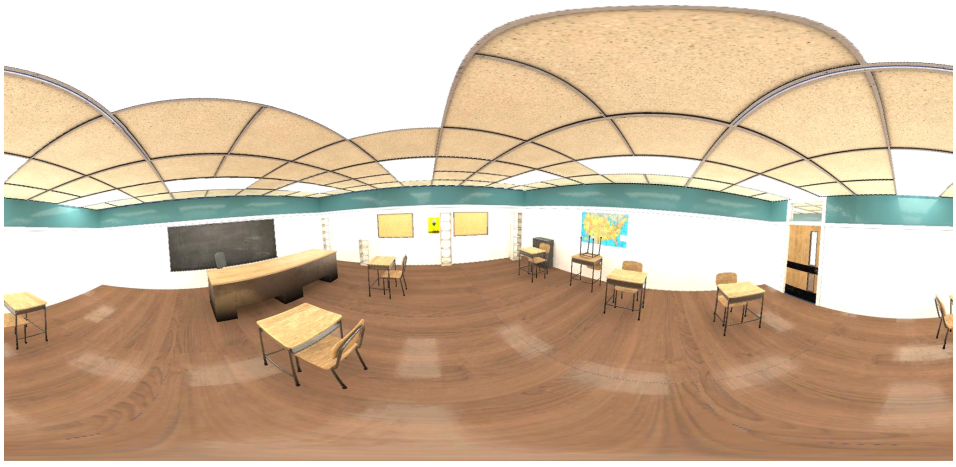


[그림 A.1-5] 구역5의 전방향 영상

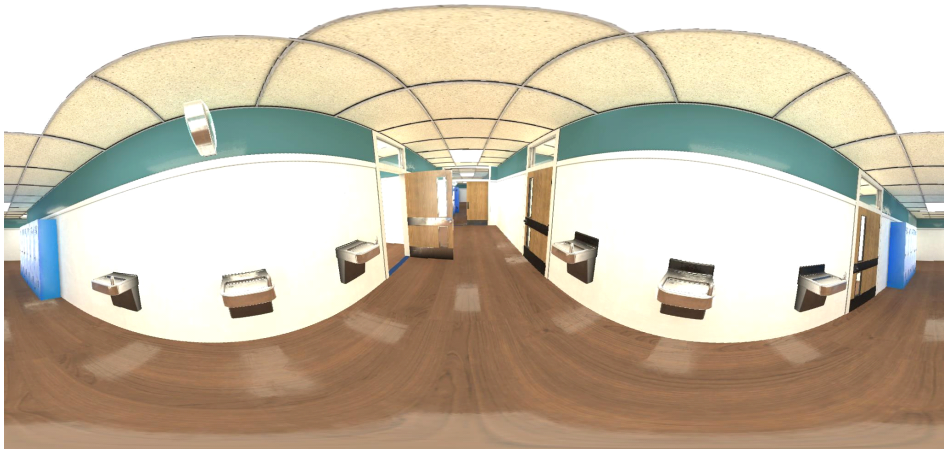


[그림 A.1-6] 구역6의 전방향 영상

[그림 A.1-7] 구역7의 전방향 영상



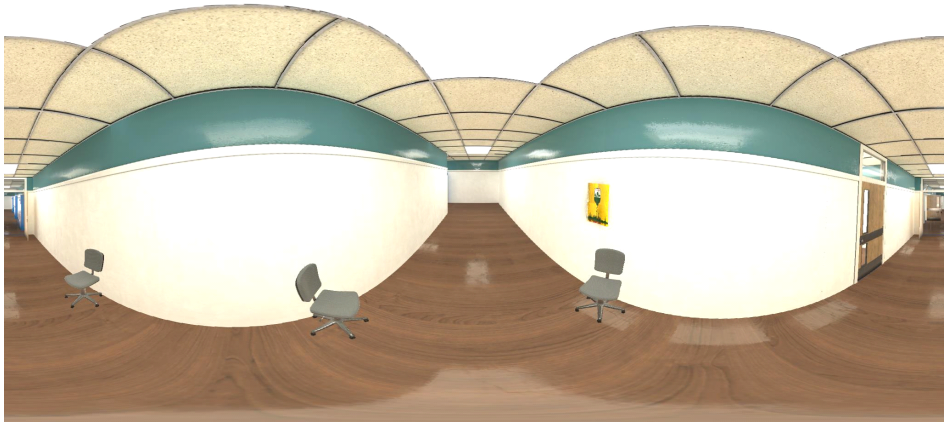
[그림 A.1-8] 구역8의 전방향 영상



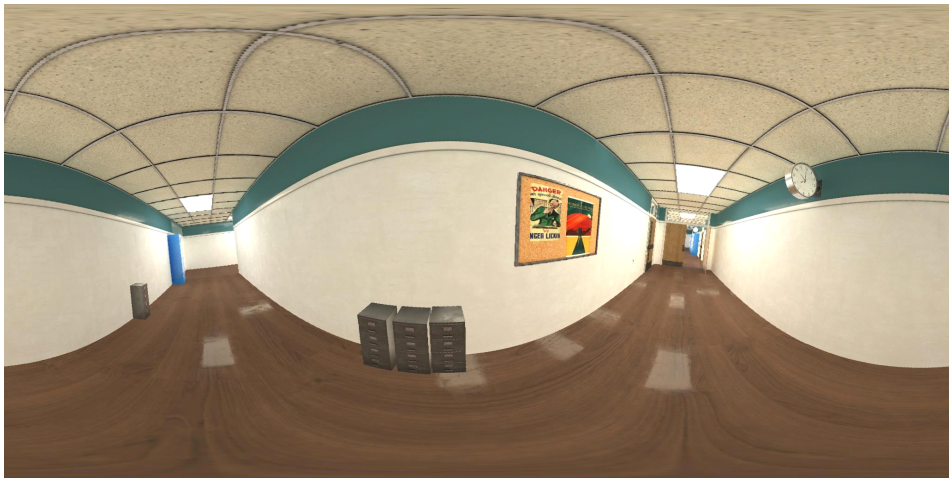
[그림 A.1-9] 구역9의 전방향 영상

[그림 A.1-10] 구역10의 전방향 영상

[그림 A.1-11] 구역11의 전방향 영상



[그림 A.1-12] 구역12의 전방향 영상



[그림 A.1-13] 구역13의 전방향 영상

[그림 A.1-14] 구역14의 전방향 영상

[그림 A.1-15] 구역15의 전방향 영상

A.2 모든 경우의 Confusion Matrix

[표 A.2-1] AlexNet 전이학습 유형 1의 Confusion Matrix

구역	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	103	8	11	12	2	16	28	10	1	0	0	0	1	2	0	194
2	22	194	10	11	1	4	15	9	0	1	0	0	1	1	0	269
3	7	8	229	6	0	3	5	2	0	0	1	1	2	0	0	264
4	38	25	11	227	0	14	19	9	0	4	0	0	1	0	0	348
5	6	17	2	6	271	7	4	26	0	0	1	1	1	1	0	343
6	16	10	3	13	6	222	19	11	3	0	0	1	1	2	0	307
7	44	2	6	10	1	12	178	5	1	0	0	0	0	0	0	259
8	22	10	7	12	13	16	5	224	0	0	0	0	0	1	0	310
9	2	1	1	1	1	1	3	0	185	3	8	1	3	51	7	268
10	3	3	3	1	0	1	1	0	10	192	26	3	8	15	7	273
11	3	5	7	1	1	0	4	0	25	5	177	16	11	12	31	298
12	9	6	3	0	2	1	8	2	3	19	6	227	31	1	0	318
13	16	8	6	0	2	1	6	2	7	60	22	46	222	20	0	418
14	9	3	0	0	0	1	5	0	43	11	5	4	18	166	16	281
15	0	0	1	0	0	1	0	0	22	5	54	0	0	28	239	350
	300	300	300	300	300	300	300	300	300	300	300	300	300	300	300	4500

[표 A.2-2] AlexNet 전이학습 유형 3의 Confusion Matrix

구역	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	156	5	6	6	3	8	47	6	0	0	1	0	0	0	0	238
2	19	225	17	1	1	6	17	11	0	1	0	1	1	1	1	302
3	2	2	248	5	0	0	1	4	0	1	5	0	1	0	0	269
4	56	17	6	273	7	19	25	17	0	5	2	1	2	1	0	431
5	2	10	1	2	285	5	4	14	0	0	0	0	2	0	0	325
6	5	3	1	0	1	242	10	11	1	5	1	0	0	1	0	281
7	20	4	4	3	0	2	171	3	0	0	1	0	0	0	0	208
8	13	11	8	10	2	12	10	233	0	2	1	0	1	0	0	303
9	2	2	1	0	0	1	2	0	253	2	8	0	10	26	0	307
10	2	7	0	0	0	1	1	1	10	227	36	3	18	11	2	319
11	5	4	5	0	0	0	1	0	3	7	197	7	8	5	15	257
12	11	7	1	0	0	2	9	0	2	26	12	272	75	6	0	423
13	5	2	2	0	0	1	0	0	5	17	4	14	169	11	0	230
14	1	1	0	0	0	0	1	0	14	3	5	2	13	211	7	258
15	1	0	0	0	1	1	1	0	12	4	27	0	0	27	275	349
	300	300	300	300	300	300	300	300	300	300	300	300	300	300	300	4500

[표 A.2-3] MobileNet V2 전이학습 유형 1의 Confusion Matrix

구역	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	166	11	6	9	6	8	36	7	0	1	0	1	0	1	0	252
2	9	211	3	3	0	11	14	8	2	1	2	1	0	0	0	265
3	4	1	267	9	0	1	1	2	0	0	1	0	0	0	0	286
4	41	22	4	262	0	17	17	6	0	0	0	1	1	0	0	371
5	2	11	2	3	292	2	2	17	0	1	2	2	5	0	0	341
6	6	2	1	1	0	243	8	8	0	1	1	0	0	0	0	271
7	35	8	3	6	1	6	195	6	1	0	0	0	1	1	0	263
8	9	12	5	6	1	8	9	245	0	0	0	0	1	0	0	296
9	1	0	0	0	0	1	4	0	246	2	6	0	5	24	0	289
10	0	4	1	1	0	0	0	0	7	218	20	2	7	12	0	272
11	8	2	3	0	0	1	3	0	13	9	222	7	14	3	7	292
12	15	12	2	0	0	1	7	0	2	22	9	269	61	5	0	405
13	3	3	2	0	0	1	2	0	4	31	7	17	195	15	0	280
14	0	1	0	0	0	0	2	1	14	12	4	0	10	209	8	261
15	1	0	1	0	0	0	0	0	11	2	26	0	0	30	285	356
	300	300	300	300	300	300	300	300	300	300	300	300	300	300	300	4500

[표 A.2-4] MobileNet V2 전이학습 유형 3의 Confusion Matrix

구역	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	159	10	8	16	0	8	39	16	0	0	0	0	0	1	0	257
2	18	210	4	4	0	4	9	10	0	0	1	0	1	0	0	261
3	10	12	252	6	0	6	5	4	1	0	2	1	3	0	1	303
4	25	13	3	235	0	9	2	7	0	1	0	0	0	0	0	295
5	2	7	3	7	283	2	2	13	0	1	1	0	2	1	0	324
6	17	7	1	10	0	247	13	16	2	2	0	0	3	0	0	318
7	32	9	9	11	3	7	198	3	1	0	1	1	0	2	1	278
8	9	8	4	7	8	11	4	224	1	1	0	1	0	0	0	278
9	0	4	0	0	1	1	3	0	221	0	3	3	3	30	1	270
10	1	8	0	3	1	1	1	0	15	222	20	7	25	20	2	326
11	2	1	8	0	0	0	5	1	23	7	209	14	11	8	24	313
12	11	4	7	0	1	0	9	4	1	17	11	246	35	5	0	351
13	9	4	1	0	3	1	4	1	5	36	14	24	208	12	1	323
14	4	3	0	1	0	2	5	1	26	11	4	3	9	205	6	280
15	1	0	0	0	0	1	1	0	4	2	34	0	0	16	264	323
	300	300	300	300	300	300	300	300	300	300	300	300	300	300	300	4500

[표 A.2-5] Inception-ResNet V2 전이학습 유형 1의 Confusion Matrix

구역	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	170	8	4	10	2	13	32	17	1	0	0	0	0	0	0	257
2	21	219	2	3	0	5	15	3	0	0	0	0	0	0	0	268
3	4	0	267	4	0	0	1	3	0	1	2	1	5	0	1	289
4	21	21	9	259	1	14	10	12	0	2	0	0	0	0	0	349
5	9	11	2	3	282	3	5	15	1	6	3	2	9	1	0	352
6	6	13	1	1	3	249	11	15	0	0	1	1	0	1	0	302
7	26	4	5	3	2	6	205	6	0	0	1	0	0	0	0	258
8	22	6	1	14	9	8	5	224	0	0	0	0	0	0	0	289
9	2	0	0	0	0	0	2	0	226	1	1	0	6	37	0	275
10	0	1	0	0	0	0	0	0	10	221	19	4	16	6	3	280
11	4	1	5	0	0	0	3	0	26	7	221	11	10	0	12	300
12	5	6	3	1	0	2	5	4	1	12	5	260	37	7	0	348
13	5	9	1	0	1	0	4	1	8	47	9	20	207	11	0	323
14	4	1	0	1	0	0	2	0	19	2	7	1	10	218	5	270
15	1	0	0	1	0	0	0	0	8	1	31	0	0	19	279	340
	300	300	300	300	300	300	300	300	300	300	300	300	300	300	300	4500

[표 A.2-6] Inception-ResNet V2 전이학습 유형 3의 Confusion Matrix

구역	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	212	9	2	7	1	5	26	2	0	0	0	0	1	0	0	265
2	15	245	0	0	0	0	5	10	0	0	1	0	1	0	0	277
3	0	1	285	1	0	0	0	0	0	0	5	0	3	0	0	295
4	13	7	0	277	1	11	3	6	0	0	0	0	0	0	0	318
5	8	8	0	4	292	2	1	9	0	0	1	0	1	0	0	326
6	5	3	1	2	2	275	7	9	0	0	0	0	0	1	0	305
7	26	8	1	1	2	4	243	5	0	0	1	0	0	0	0	291
8	7	6	4	7	2	3	4	258	0	0	0	0	0	0	0	291
9	0	2	0	0	0	0	1	0	264	1	1	0	1	14	0	284
10	0	3	0	0	0	0	0	0	10	256	15	4	15	6	0	309
11	4	0	2	0	0	0	0	0	7	4	259	6	2	3	3	290
12	3	0	1	0	0	0	2	1	0	9	0	281	22	3	0	322
13	2	5	3	1	0	0	4	0	4	27	6	9	248	10	0	319
14	5	3	0	0	0	0	4	0	14	1	4	0	6	256	4	297
15	0	0	1	0	0	0	0	0	1	2	7	0	0	7	293	311
	300	300	300	300	300	300	300	300	300	300	300	300	300	300	300	4500

Abstract

A Study on Image Based Deep Convolutional Neural Network for Indoor Positioning

- Method of Data Refinement with Omnidirectional Image and Comparison of Transfer Learning Methodologies -

Gwangjoong Kim

Department of Civil and Environmental Engineering

The Graduate School

Seoul National University

Location-based services can be used for the introduction of Internet of Things (IoT) in various industries, the spread of high-speed communication networks, real-time location tracking and traffic information, location-based search and advertising and marketing. In

location-based services, location estimation technology plays an essential role as a basic data base for providing appropriate services. Therefore, accurate positioning technology is needed to improve the quality of location based services. However, global satellite navigation systems, which are commonly used position estimation techniques, are relatively inaccurate at indoor space and require different technologies. Among such technologies, the method of estimating the indoor position using the sensor mounted on the smartphone is superior in terms of accessibility due to the spread of the smartphone. In particular, the method of classifying the images acquired by the smartphone camera and estimating the position is economical because it requires no additional equipment installation, and is independent on the internet, so that it can operate stably. Recently, the image classification system using the convolutional neural network has been greatly developed, and the indoor positioning system using the convolutional neural network has received attention.

However, in order to fully train the convolution neural network with good performance from the scratch, a large amount of training data, a calculation resource and a calculation time are required. Therefore, in most cases, transfer learning which uses pre-trained convolution neural networks have been used. The following factors should be considered in order to learn the convolutional neural network using transfer learning.

First, we need to determine the type of deep convolution neural network to use for transfer learning. Since AlexNet developed, deep convolutional neural networks developed in terms of efficiency and accuracy using several techniques. Therefore, it is necessary to compare the performance of the convolutional neural networks in indoor positioning for image classification. Next, selecting the transfer

learning type is needed. Transfer learning can be classified into four types according to the degree of similarity between target data and pre-training data and the number of target data. Therefore, it is necessary to set the transfer learning type and to compare the performance between the set transfer learning types in using the deep convolutional neural network. Even using the transfer learning, the average number of training images per zone was more than 1,000 in the previous studies. Therefore, it is necessary to devise a method for efficiently acquiring a large amount of images. Finally, a method of removing data that leads to poor accuracy should be studied. When acquiring images in large quantities, images that adversely affect neural network learning may be included. Research to remove them should be backed up in the way of obtaining a large amount of images.

This study compared the performances of types of neural networks and types of transfer learning in indoor positioning by image classification through deep convolutional neural network. And I proposed a method of using omnidirectional image to efficiently acquire a large amount of images. In addition, I proposed a method to remove images that adversely affected on training when omnidirectional images were used.

In this study, I used AlexNet, MobilNet V2, and Inception-ResNet V2 among the exist neural networks. The above-mentioned neural network was known to be a neural network for image classification of ImageNet, which is an image database used in ILSVRC (ImageNet Large Scale Visual Recognition Competition), and it is known that ImageNet image classification accuracy is higher in order of Inception-ResNet V2, MobileNet V2 and AlexNet.

Since the experiments in this study were performed in the virtual

space, I explained how to construct the virtual space for the experiment first. Based on the assumption of the actual situation, the common points and characteristics were given to each place, and the orientation and the path of the omnidirectional camera were set.

In this paper, I proposed a method to generate multiple pinhole camera models from an omnidirectional image in which color information is mapped to a unit sphere, and then divide them into perspective projection images. In the proposed method, 30 perspective images were generated from one omnidirectional image.

Using the generated perspective projection images, AlexNet, MobilNet V2, and Inception-ResNet V2 transfer learned from the pre-trained from ImageNet images. In order to analyze for each type of transfer learning, the training was carried out in six cases by learning the type 1 which only trains the fully connected layer for each neural network, and the type 3 which trains the part of the convolutional layers partly.

Data removal was performed using entropy to calculate the amount of information in the image and neural network learned with the image not removed. The entropy criterion of the image to be removed was increased until the frequency of the right image classification becomes larger than the frequency of the misclassification.

As a result of the experiment, it was confirmed that using the neural network with high image classification accuracy of ImageNet is advantageous for the indoor positioning in terms of accuracy, because the indoor positioning accuracy is high in order of the image classification accuracy of ImageNet. In addition, the accuracy was 6.12% higher in transfer learning type 3 than in transfer learning type 1, so that type 3 is more advantageous for indoor positioning

than type 1. Also, I confirmed that the neural network learned after data removal improved the average accuracy by 1.99% compared with the neural network learned by the data that was not removed.

As a result of this study, it is more advantageous to use the advanced neural network than to use the neural network initially developed and to train the layer for feature extraction than to learn only the fully connected layer. In addition, the proposed method of generating a large-scale perspective projection image using the omnidirectional image is worthy of being able to solve the problem of data quantity shortage in the learning of the convolutional neural network. And the method of eliminating the image which adversely affects the learning can be used universally for the method of obtaining a large amount of images. The results of this study can be applied to the indoor positioning studies using the convolutional neural network and the image based convolutional neural network using the transfer learning can be applied as one of the means of indoor positioning.

keywords : Indoor Positioning, Deep Convolutional Neural Network, Omnidirectional Image, Data Refinement

Student Number : 2017-22313